# Systems Based Analysis of Gastric Cancer Myofibroblasts

**Helen Louise Smith**

**Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy**

**September 2012**

# Systems Based Analysis of Gastric Cancer Myofibroblasts

Helen Louise Smith

## Abstract

The tissue microenvironment plays an important role in tumour development and the regulation of cancer cell migration. Tumours have been referred to as 'wounds that do not heal', as cancer cells exploit the normal wound response to aid tumour development and metastasis. During this process cancer cells secrete a range of factors that influence the function of cells within the surrounding tissue. Myofibroblasts are the predominant cell-type in the cancer microenvironment, where they surround the developing tumour. Cancer cells have the ability to re-program these myofibroblasts in order to provide a range of factors that are required to promote the growth, proliferation and migration of cancer cells.

In this study a combination of bioinformatic and statistical approaches were used to define the range of genes and biological pathways that are differentially regulated in myofibroblasts derived from the immediate tumour microenvironment (CAMs) or from adjacent gastric tissue (ATMs). Comparison of gene expression profiles between CAMs, ATMs and normal gastric myofibroblasts (ANs) identified molecular signatures and biological processes that are changed in CAMs and/or ATMs. By correlating patient prognosis scores with CAM gene expression profiles it was possible to show that patient with poor prognosis segregate into two distinct populations with distinct gene expression profiles. Pathway enrichment and multivariate correspondence analysis, show that CAMs and ATMs both undergo metabolic reprogramming to induce a variation of the 'Reverse Warburg effect' in

which myofibroblasts exhibit increased levels of fatty acid β-oxidation and enhanced production of acetyl-CoA and ketone body biosynthesis, along with the up-regulation of mono-carboxylate transporters that are required to facilitate the import of fatty acids and the export of ketone bodies. This data defines prognosis specific expression profiles, thereby providing new insight into the molecular processes that drive important paracrine communication networks during the development of gastric tumours. As such, this data provides a resource for future experimental studies and biomarker development.

## Table of Contents

# Appendices

List of supplementary files provided on CD:

*Supplementary chapters:*

Supplementary Chapter 1

*Supplementary excel files:*

Chapter 3 - Supplementary excel files 3.1- 3.24

Chapter 4 - Supplementary excel files 4.1-4.9

Chapter 5 - Supplementary excel files 5.1-5.11

*Supplementary R scripts:*

Supplementary script 1 (Reactome)

Supplementary script 2 (AffyQualityMetrics)

Supplementary scripts 3 (Reactome pathway mapping)

Supplementary scripts 4 (Statistical tests and odds ratios)

Supplementary scripts 5 (Matrix formation)

Supplementary scripts 6 (MVA analysis, similarity of genes)

Supplementary scripts 7 (MVA analysis, similarity of pathways)

Supplementary scripts 8 (Correspondence analysis)

Supplementary scripts 9 (Correlation analysis, CAM vs. AN)

Supplementary scripts 10 (Correlation analysis, CAM vs. ATM)

Supplementary scripts 11 (Correlation analysis, ATM vs. AN)

Supplementary scripts 12 (T-test, 'good' and 'bad' patient prognosis)

## Courses and Conferences

**Courses attended**

Personal animal licence 2007

First year skills workshop, Liverpool 2008

Second year skills workshop, Bournemouth 2009

Presentation skills workshop 2010

**Conferences attended**

Cancer research UK symposium, Liverpool 2010 (Presented poster)

Cancer research UK symposium, Liverpool 2011 and 2012 (Supervisor presented our research)

Network Biology Symposium, Chelsea, 2010.

## Acknowledgements

Firstly, I would like to thank my supervisor Chris Sanderson for his continued support, guidance and enthusiasm throughout the entirety my project. All past and present group members Julie, Russell, Jon, Rob, Kelly, Emily, Hanna, Dave and Amy, for making coming into the office (and the AJ) a much more pleasant experience. An additional thank you to Russell for introducing me to the wonderful world of bioinformatics and Jithesh and Richard for their continued statistical support. To all the students in my year (Chris, Hanna, Sarah, Seb, Michelle and Stuart) for the much-needed extracurricular distractions, I wish you every success within your future careers.

A special thank you to my family for their continued encouragement, especially to my mum and dad who always ensured I strived to reach my full potential. To Alfie, who was a mere 4 months old when I started my PhD, has been the most perfectly behaved little boy. Thank you for always sleeping through the night so I have the strength to come into work, for your calm temperament and heart-warming smile, after all, everything I do is for you. Last but certainly not least, to my husband Danny who has always supported me, financially and emotionally throughout my research. You have always had an unwavering belief in me, for which I am forever grateful.

# 1 Chapter One: Introduction

## 1.1 The Aetiology of Gastric Cancer

According to Cancer Research UKs mortality statistics (2010) stomach cancer remains one of the top ten causes of cancer related deaths in the UK and incidence in far eastern countries (Japan and Korea) is higher (Jemal et al., 2010). Unfortunately, post diagnosis survival rates are very poor for patients with gastric cancer. This is primarily due to the fact that patients tend not to be symptomatic until the tumour is in an advanced stage of development, in many cases tumours may already have spread to lymph nodes and secondary tissues/organs before diagnosis (Matsukuma et al., 1996). For this reason we need to develop a better understanding of the molecular and functional changes that drive the development and spread of gastric tumours. Better insight into these processes could provide new options for developing better therapeutic strategies and the identification of possible diagnostic signatures or trends to improve detection or prognostic predictions. Risk factors relating to gastric cancer include smoking, poor diet, excessive alcohol consumption, helicobacter pylori (*H.pylori)* infection and a range of hereditary genetic factors (Fuchs and Mayer, 1995; Li et al., 2012; You et al., 1998). *H.pylori* infection induces chronic inflammation leading to gastric atrophy and neoplasm, which may then lead to the development of an adenocarcinoma (Atherton, 2006; Kusters et al., 2006). A minority of gastric cancers may be lymphomas; cancer of the white blood cells (Parsonnet et al., 1994). Gastric atrophy can result from chronic inflammation of gastric tissues, which may lead to loss of acid producing parietal cells, which in turn increases gastrin production, thereby

stimulating epithelial cell proliferation (Sipponen et al., 1985; Vannella et al., 2012). Gland preservation is considered the key event to prevent a cancerous outcome (Correa et al., 2006). The constant regenerative response of chronic gastritis can lead to intestinal metaplasia, which is the change of a differentiated cell into another cell-type, in this case in response to gastritis the regenerative process replaces cells with columnar mucosa like intestinal cells. Intestinal metaplasia predisposes the patient to a much higher risk of developing gastric cancer if accompanied by dysplasia (You et al., 1998). Dysplasia is characterized by nuclear abnormalities, cytoplasmic abnormalities, and increased rate of cellular multiplication (LAUREN, 1965), it is a state of abnormal differentiation, which eventually leads to adenocarcinoma (Lauwers and Riddell, 1999; Oehlert et al., 1979). Adenocarcinoma is the name of a malignant epithelial tumour, originating within glandular tissue. Histologically, adenocarcinoma can be separated into intestinal or diffuse types; intestinal adenocarcinoma resembles tubular type structures, whilst diffuse adenocarcinoma secretes mucus resulting in large pools of mucus/colloid. This histological classification of the stages of tumour development is commonly known as the Lauren classification (Shibata et al., 2001). A diagram of the multi-step gastric carcinogenesis pathway, which is believed to be initiated by *H.pylori* is shown in Figure 1.1.

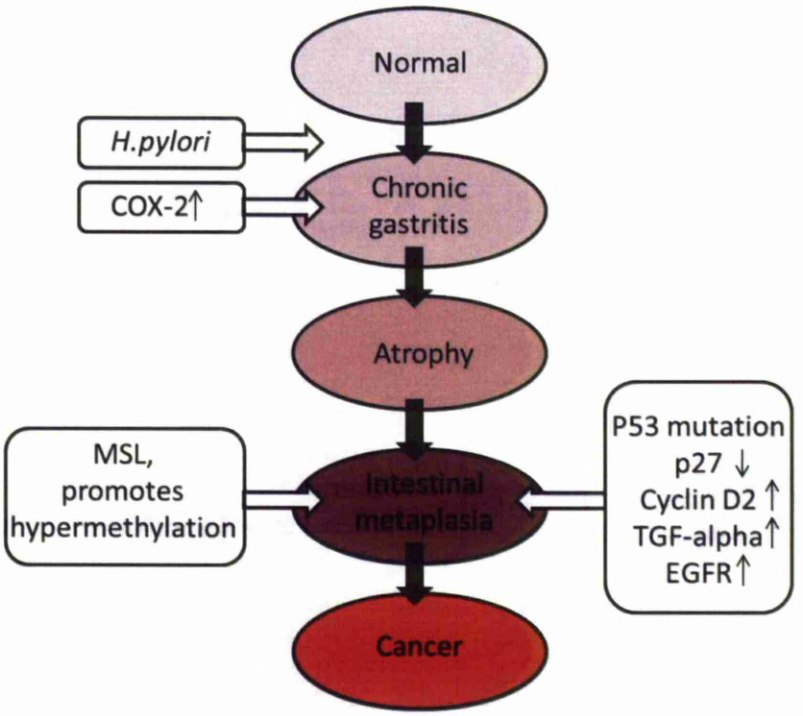



**Figure 1.1 The multi-step gastric carcinogenesis pathway, believed to be initiated by *H.pylori*. COX-2, cyclo-oxygenase -2; EGFR, epidermal growth factor receptor; MSL, microsatellite instability; TGF, transforming growth factor. Figure adapted from (Leung et al 2002).**

## 1.2 Inflammation and Cancer

Inflammation is caused by the body's response to infection or irritation, resulting in an increased repair response. However, long-term chronic inflammation can lead to initiation and development of cancer. The apparent synergy between these two processes may not be surprising, as both processes tend to suppress cell death (apoptosis) and increase cell growth and angiogenesis.

In general, inflammation increases the infiltration of inflammatory cells, resulting in increased local production of free radical species such as nitric oxide and super

oxides. Free radicals are produced to combat microbial infection, this process can also damage DNA within host cells (Hussain et al., 2003; Rakoff-Nahoum, 2006). Leading to increased apoptosis, as cells attempt to protect the body against replication of potentially dangerous changes in DNA structure or sequence. The loss of cells leads to a need for tissue repair and regeneration, which is facilitated by further inflammation signals (Chen et al., 2003), including the secretion of pro-inflammatory interleukins (ILs) and cyclooxygenase-2 (COX-2) enzymes. An increase in IL-8 causes the release of IL-1β and Tumour Necrosis Factor alpha (TNF-α) from immune cells, stimulating the expression of COX enzymes, thereby stimulating production of prostaglandins, which further enhance the inflammatory response, leading to a further increase in nitric oxide production and a general increase in tissue acidity (Beales and Calam, 1998). Significantly, COX-inhibitors have been shown to decrease the tumour severity (Akre et al., 2001) and interference in pro-inflammatory responses remains an active area of anti-cancer research.

The importance of infection in initiating inflammation and tumour development was first demonstrated in 1985 using the Rous sarcoma virus (Dolberg et al., 1985). Currently it is estimated that around 15% of all cancers are initiated by microbial infection, with the prominent cause of gastric cancer being chronic infection with the bacterium *Helicobacter pylori* (*H. pylori*). *H. pylori* induces inflammation of the gastric mucosa. In most cases, inflammation remains superficial and does not develop into atrophic gastritis or gastric cancer. People most susceptible to atrophic gastritis are infected with the *H. pylori cag*$A^+$ strain, which increases epithelial cell proliferation through gastrin secretion (Wang et al., 2000). In addition, *H.pylori*

inhibits the p53 gene through AKT, as p53 roles include protecting the cell from DNA damage this is a potential way *H.pylori* could increase the risk of gastric cancer (Wei et al., 2010).

## 1.3 Common factors in cancer progression

Over the years, common hallmarks to all cancers have become apparent; these include proliferative signalling, replicative immortality, the avoidance of growth suppressors, resistance to cell death, angiogenesis, local invasion and metastasis (Figure 1.2) (Hanahan and Weinberg, 2011; Hanahan and Weinberg, 2000). Studies have been carried out to identify cell lineage-specific genes that are essential for cancer development (Luo et al., 2008). In general, genes may become differentially regulated due to acquisition of somatic mutations or alterations in epigenetic regulation.

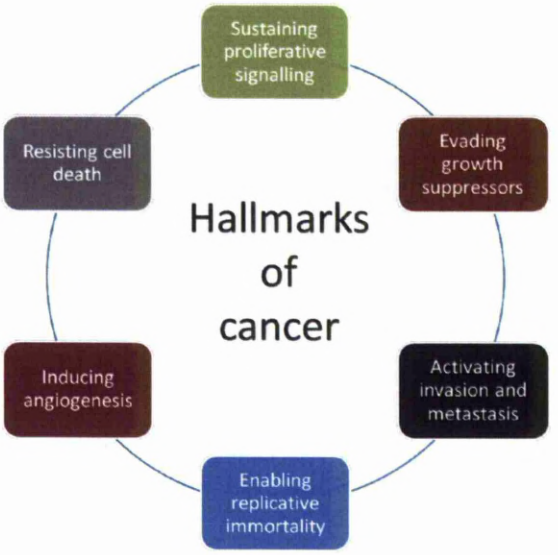



**Figure 1.2. Six Hallmarks of cancer. Figure adapted from (Hanahan and Weinberg, 2011).**

Gene mutations may occur naturally during cell duplication or they may be induced by exposure of oncogenic agents. Driver mutations occur in cancer cells and may provide the cell with a growth advantage, whereas passenger mutations are phenotypically neutral (Greenman et al., 2007). A mutation may also be deleterious, were the expression of the gene is reduced or altered as a result of in-frame mutations that do not prevent gene expression. The exact position of an acquired mutation will have differential effects on gene expression, depending on its locality to the last splice site junction and the ability to trigger transcript degradation via the nonsense-mediated decay (NMD) pathway. The resulting level of expression of an in-frame mutation, determines the observed phenotype (Zhong et al., 2009). Somatic driver mutations have been shown to cause de-regulation of signalling pathways, avoidance of cell death and increased proliferation and invasion (Davies and Samuels, 2010). Reduced activity of the proto-oncogene KRas, leads to a reduction in the cyclin dependent inhibitor p21, which in turn increases the potential for metastasis (Alon et al., 1987).

Epigenetics is the regulation of gene expression though modifications in histone function, chromatin structure, DNA modification or the action of small non-coding microRNAs, rather than alteration in the genomic sequence. Tumour suppressors are often epigenetically regulated in cancers. For example, the tumour suppressor retinoblastoma (RB) protein is known to epigenetically regulate expression of the E2F family of transcription factors, thereby inhibiting DNA replication by recruiting histone deacetylases to promoter regions. Cyclin dependent kinases phosphorylate RB, thus preventing inhibition of gene transcription. In many cancers, mutations/

loss of RB and cyclin dependent regulatory proteins, result in an over-expression of cyclin dependent kinases, resulting in a loss of senescence (Berdasco and Esteller, 2010; Sherr and McCormick, 2002).

The Ras proteins is involved in increasing cell proliferation, differentiation and inhibition of apoptosis. As with p53, Ras GTPases are often mutated in many cancers, however, this does not appear to be the case in many gastric cancers, where RAS expression appears to be de-regulated by epigenetic effects. The RAS activator like-1 gene (GAP1 like) is hyper-methylated and its expression is reduced, consequently increasing the concentration of RAS, resulting in an increase in mitogen activated protein kinase and stimulating tumourogenesis. Also, Ras activator like-1 protein expression can be restored and proliferation reduced by inhibitors of histone-deacetylase (Seto et al., 2011). The promoter region of E-cadherin has also been shown to be hyper-methylated in many cancers, causing a decrease in E-cadherin expression and increased tumour invasion (Kanazawa et al., 2002a).

### 1.3.1 Tumour suppressors and un-controlled cellular growth

Tumour suppressors are commonly de-regulated during cancer. The tumour suppressor p53 detects stress and promotes expression of cell-cycle inhibitors in order to induce senescence or apoptosis. Genetic mutations within the RB and p53 pathways provide a proliferative advantage in the progression of cancer, up to 50% of cancers have mutations in the p53 gene, enabling cells to bypass senescence (Sherr and McCormick, 2002). In normal circumstances, cells experiencing constant stress, through lack of nutrients, or detection of DNA damage would undergo

apoptosis. The BCL-2 family of proteins include pro and anti-apoptotic regulators which are often defective in tumours, therefore BCL-2 proteins and their binding partners are possible targets for cancer therapy (Adams and Cory, 2007).

Autophagy is a stress response that causes cells to recycle less essential intracellular components to aid cell survival. Tumour cells can take advantage of this process in order to remain in a low energy state when nutrients are sparse, or by suppressing the autophagy inducer Beclin-1 (Sinha and Levine, 2008; White and DiPaola, 2009).

### 1.3.2 Angiogenesis

Proliferating cells quickly use up nutrients within their immediate microenvironment; therefore they have evolved mechanisms of increasing local resources. Firstly, by inducing angiogenesis (Folkman et al., 1971) to obtain nutrients via increased blood supply and secondly by secreting factors that induce neighbouring cells to provide a selection of metabolic substrates (Mazure et al., 1996). Importantly, angiogenesis has been detected in pre-malignant tumours therefore indicting it is an early change in cancer development and a possible target for therapeutic intervention (Raica et al., 2009).

Hypoxia has also been suggested to be the trigger for another trait associated with cancers, known as the Warburg effect (Warburg, 1956), or aerobic glycolysis is the process of cells converting glucose to lactate when oxygen is available in the micro-environment (Pennacchietti et al., 2003). This process of converting pyruvate to lactate usually takes place under low oxygen conditions, however, in tumour cells this occurs in the presence of oxygen (Christofk et al., 2008). Therefore, Warburg suggested that the production of lactate in high oxygen conditions may be due to a

defect in the mitochondria. In fact no such mitochondrial defect was found, instead cancer cells appear to produce high levels of lactate dehydrogenase and aerobic glycolysis can confer a selective growth advantage, in low oxygen conditions, as glucose is broken down aerobically to produce macromolecules and other essential factors that facilitate cell proliferation (Fantin et al., 2006).

### 1.3.3 Invasion and metastasis

Metastasis is one of the hallmarks of late-stage malignancy. Prior to migration, tumour cells induce degradation of the basement membrane, after which they can travel through the local tissue, before entering the circulatory system (intravasion). Extravasion occurs when circulating cells leave the vascular system and enter distal tissues, where they form new distal tumours (Hanahan and Weinberg, 2011). Formation of new tumours at metastatic sites is dependent on the microenvironment that exists in different organs (Pauli and Lee, 1988). In general, epithelial cells are very structured, uniformly aligned and have tight regular cell-cell adhesions, which inhibit cell movement (Lee et al., 2006). In order to become mobile and invasive, they have to be transformed from an epithelial to a mesenchymal phenotype. This epithelial-mesenchymal-transition (EMT) is a reversible morphological and physiological change that is seen when an epithelial cell invades the extracellular matrix. The reverse process is referred to as the mesenchymal epithelial transition (MET), which generally results in the formation of new colonies, either in the same tissue, or in distal organs.

During the EMT process, epithelial cells lose cellular polarity and adopt a characteristic spindle shape. In addition, cytoskeletal keratin is replaced with

vimentin, expression of cell-cell adhesion molecules are reduced and cells display increased motility (Klymkowsky and Savagner, 2009; Lee et al., 2006). Initiation of EMT is thought to be stimulated through an array of pathways including, activation of tyrosine kinase receptors, which participate in the Transforming Growth Factor beta (TGF-β), Epithelial Growth Factor (EGF), Fibroblast Growth Factor (FGF), Insulin-like Growth Factor (IGF), and the NFκB signalling pathway. Increased expression of E-cadherin directly inhibits EMT and metastatic behaviour in cancer (Kanazawa et al., 2002b). The important role of E-cadherins in gastric cancer was first confirmed by the identification of inactivating mutations and premature stop-codons in the E-cadherin gene, CDH1 (Guilford et al., 1998). Since then E-cadherin transcriptional regulators Snail, SIP1 and Twist have all been shown to be frequently inhibited or down-regulated in gastric cancer (Rosivatz et al., 2002) and the E-cadherin promoter has also been shown to be methylated in gastric cancer (Kanazawa et al., 2002a). Interestingly, stromal myofibroblasts are believed to be the key cell type mediating the EMT transition (Trujillo et al., 2010).

## 1.4 The cancer microenvironment

There is now convincing evidence to show that the tissue microenvironment plays an important role in tumour development and the regulation of cancer cell migration. In 1988, Paget first proposed the seed and soil concept to explain the role that the tissue microenvironment plays in tumour development (Paget, 1989). In this analogy, Paget likens the relationship of a cancer cell and its microenvironment to that of a seed planted in different locations. Whether the seed or cancer cell grows depends on whether the environmental conditions are

favourable or not. It is now clear that this analogy is a 'good' reflection of events that regulate the initial stages of tumour development and determine preferential sites of metastasis for different tumours. We now know that cancer cells have the ability to re-program neighbouring cells within the surrounding tissue in order to provide a range of factors that are required to promote the growth, proliferation and subsequent migration of cancer cells (Witz, 2009).

Although all tissues are composed of a complex mix of different cell types, there are key factors that contribute to the development of tumours in several different tissues. These include factors that affect the integrity of the extracellular matrix, immune cell infiltration, and the role of stromal myofibroblasts. The extracellular matrix provides structural support and acts as a reservoir of nutrients, growth factors and other regulatory factors; it is comprised of a basement membrane, blood capillaries and matrix proteins including collagen, fibronectin and elastin, which are secreted by cells in the extracellular matrix including fibroblasts, adipocytes and immune cells (Bhowmick et al., 2004b; Dvorak, 1986). Dense growth of fibrotic and connective tissue is often seen surrounding tumour cells and this phenotype is referred to as desmoplasia (Willis R, 1961).

Tumours have been referred to as 'wounds that do not heal', as the normal wound response is exploited by cancer cells to aid tumour progression. Normally in response to injury, blood vessels are damaged and release fibrin and fibronectin, which cause an increase in the migration of inflammatory cells and fibroblasts, and promote the formation of new capillaries. Fibroblasts secrete extracellular materials to reconstruct the extracellular matrix. When the wound response is complete,

fibroblasts migrate away from the wound and a scar is formed. Cancer cells exploit this response by secreting a vascular permeability factor, which causes the blood capillaries to become permeable, allowing fibrin and fibronectin to diffuse into the extracellular matrix. Cancer cells also secrete growth factors that increase fibroblast migration and proliferation. Increased levels of fibrin and fibronectin also promote angiogenesis and migration of fibroblasts and immune cells. However, in tumours, the concentration of fibronectin and fibrin does not decrease, as would normally happen in wound healing (Dvorak, 1986). Analysis of twenty solid malignant tumours showed that macrophages and lymphocytes where found mainly in stroma surrounding the tumour cells rather than within tumours. Interestingly, stomach cancer stroma was found to have one of the highest populations of macrophages and lymphocytes, which is believed to be due to the prevalence of a non-specific inflammatory response in this tissue (Svennevig and Svaar, 1979). While there is clear immune response in cancer (Vose et al., 1977) it has been shown that normal immune responses are suppressed within the tumour microenvironment (Vose and Moore, 1979). Also, although sub-population of lymphocytes have been shown to inhibit the cytotoxic response, identification of suppressor factors has proved problematic, with several groups proposing conflicting findings (Vose and Moore, 1979; Yu et al., 1977). Significantly, treatment of peripheral lymphocytes from cancer patients, with a lethal macrophage agent carrageen was found to increase the cytotoxicity of the lymphocytes, indicating that macrophages may be the drivers of immune-suppression within the cancer microenvironment (Quan and Burtin, 1978). However, the actual affect that immune infiltration may play in the

orchestration of tumour development in different tissues or individual patients remains to be defined.

### 1.4.1 Myofibroblasts

In general, fibroblasts are quiescent cells with low motility, which are found in all normal connective tissues. They are defined by a characteristic spindle-shaped appearance and expression of high levels of vimentin. Myofibroblasts were originally defined as a subset of fibroblasts that exhibited smooth muscle like features and expressed characteristic marker proteins in addition to vimentin: these include myosin (VM-type myofibroblasts), smooth muscle actin (α-SMA) (VA-type myofibroblasts) or α-SMA and desmin (VAD-type myofibroblasts). This study focuses on the role of VA-type myofibroblasts in the development of gastric tumours. It has been shown that myofibroblasts can form from the trans-differentiation of conventional fibroblasts following stimulation with TGF-β, PDGF or the tumour niche. However, the actual origin of gastric myofibroblasts remains unclear and current evidence suggests that there may be different subclasses of myofibroblasts, possibly with different origins, however as yet no robust markers exist to define the subclasses, or the exact origins of myofibroblasts in different tissues. Interestingly, myofibroblasts, which are prominent in damaged, infected or cancerous tissue, can also be isolated from healthy gastric tissue, where they play a key role in trans-epithelial signalling, contribute to normal epithelial differentiation, enhance barrier function and modulate chloride secretion from intestinal epithelial cells (Beltinger et al., 1999; Powell et al., 1999).

There is now strong evidence that myofibroblasts also play a major role in the development and spread of tumours. In many cases myofibroblasts have been found to be the dominant cell type in the cancer microenvironment (Tuxhorn et al., 2002a), where they form 'nests' which surround the developing tumour (Samoszuk et al., 2005). Myofibroblasts were first identified in granulation tissue, which was recognized to play a key role in wound healing (Gabbiani et al., 1972; Majno, 1979). Under further investigation, it became apparent that this cell type also played a prominent role, in promoting proliferative conditions in neoplastic stroma (Zhi et al., 2010).

Although myofibroblasts have a characteristic morphology and express α-SMA there is still much that is unclear about these cells. It has been proposed that local conditions within a tumour or inflammatory niche may promote the trans-differentiation of myofibroblasts from normal fibroblasts, in response to stimulation by transforming growth factor (TGF)-β, endothelin-1 (ET-1) and IL-1 derived from epithelial cells (Chen et al., 2009; Phan, 2003; Ronnov-Jessen and Petersen, 1993; Ronnov-Jessen et al., 2002; Samoszuk et al., 2005; Tuxhorn et al., 2002a; Untergasser et al., 2005). Current theory suggests that fibroblasts firstly become 'protomyofibroblasts', which express both β and γ actins before completing the transition to active myofibroblasts, which then express α-SMA fibres (Micallef et al., 2012a). The urokinase plasmingen activator receptor (uPAR) is an extracellular protease involved in cell migration, modification of the extracellular matrix and growth factor activation. Down-regulation of uPAR is required for the differentiation of fibroblasts into 'active' myofibroblasts. As such, the loss of uPAR

is used as a marker for myofibroblast activation in some tissues (Bernstein et al., 2007).

As there is heterogeneity within populations of fibroblasts, it may not be surprising that myofibroblasts derived from these cells, will also exhibit heterogeneity in marker expression (Chauhan et al., 2003a). In addition, there is now convincing evidence to suggest that other cells can also become "activated" and display myofibroblast like characteristics. In particular, epithelial cells, bone marrow derived mesenchymal stem cells (MSCs) (Guo et al., 2008; Micallef et al., 2012b; Mishra et al., 2008) and pericytes (Pietras et al., 2008; Skalli et al., 1989) have all been identified as potential sources of activated myofibroblasts. Also, the origin of activated myofibroblasts may be tissue specific. In the liver, myofibroblast type cells expressing α-SMA are thought to be derived from both hepatic stellate cells and liver fibroblast cells.

Therefore, the origin of activated or even normal tissue myofibroblasts remains unresolved, and may well depend on the inflammatory status and the mix of stromal cells present in different tissues.

#### 1.4.1.1 Identification of myofibroblasts

At present myofibroblasts are defined only by a characteristic spindle/stellate morphology and expression of a few relatively basic markers (α-SMA, vimentin, myosin or desmin). Although, myofibroblasts share a similar phenotype with smooth muscle cells in that they both have high levels of α-SMA, the two cell types

can be clearly distinguished by the expression of additional myofibroblast markers and unique global gene expression profiles (Gan et al., 2007).

Myofibroblasts display heterogeneity between tissue and tumour type. However, some informative markers or trends in gene expression have been identified. In breast cancer associated myofibroblasts loss of expression of the transmembrane protein CD34 correlates with a more invasive phenotype. However, loss of this maker has not been detected in myofibroblasts from other types of cancer. Interestingly, CD34 was also down regulated in non-neoplastic fibroblasts surrounding epithelial tumours (Chauhan et al., 2003b). In this context it is significant to note that all myofibroblasts, whether derived from the site of a tumour (CAMs), adjacent histologically normal tissue (ATMs), or absolute normal myofibroblasts (ANM) express high levels of the marker α-SMA, thus raising the question of what is an 'activated' myofibroblast. Current markers for activated myofibroblasts permit distinctions from generic fibroblasts however; myofibroblasts tend to exhibit different functional properties, depending on their proximity to the developing tumour. As yet we do not have good markers to distinguish 'good' 'bad' or intermediate forms of myofibroblasts in tissues. As CAMs promote faster tumour growth, metastasis and have greater resistance to chemotherapy it is vital that we identify markers that distinguish the different forms of activated myofibroblasts. Equally, it has been suggested that ATMs are an intermediate between CAMs and ANMs (Hawsawi et al., 2008). However, we have no reliable marker to indicate when an ANM becomes an ATM or an ATM is about to become a far more dangerous CAM.

In 1953, it was found that although tissue surrounding the site of a tumour appeared benign, it had been changed or pre-conditioned to facilitate a high rate of reoccurrence. This phenotype was referred to as 'field cancerisation' (Slaughter et al., 1953). In one study, prostate cancer cells were combined with stromal cells from proximal or distant regions of the stroma in order to test if the 'field cancerisation' effect decreased with distance from the site of the primary tumour. Significantly, both tumour growth and angiogenesis decreased with distance (Barclay et al., 2005). In addition, in colorectal cancer patients with a high abundance of α-SMA positive myofibroblasts in the reactive-stroma had a significantly reduced chance of survival. Thus showing that myofibroblasts abundance can be used as an indicator of patient prognosis (Tsujino et al., 2007).

It is apparent that to fully understand and identify a myofibroblast subpopulation of cells, the site of carcinogenesis (or inflammation), the composition of the tumour stroma and the relative gene expression profiles of CAMs and ATMs in that specific tissue must be defined, in order to better understand the molecular mechanisms that drive proximity dependant changes, and identify better markers for myofibroblast conversion from NTMs to ATMs and CAMs.

#### 1.4.1.2 Myofibroblasts and Cancer

Reports of the role of myofibroblasts in carcinogenesis is varied; in some studies identification of myofibroblasts (by smooth muscle actin staining) has been used as a marker for patient prognosis (Fuyuhiro et al., 2010a; Fuyuhiro et al., 2010b), whilst in other studies myofibroblast recruitment/activation has been shown to correlate with increase tumour invasion, but not in the initiation of tumour growth

(de-Assis et al., 2012). Investigation into the role of myofibroblasts also varies between tissues and their role is more characterised in prostate and breast than in other tissues. In the case of prostate cancer, the tumour microenvironment is known to play a role in tumour progression and metastasis (Barron and Rowley, 2012; Bianchini et al., 2012; Giannoni et al., 2010; Shaw et al., 2009). To elucidate the mechanisms by which stromal cells drive cancer progression, Shaw et al (2009) examined the role of the sonic hedgehog pathway in paracrine signalling between the tumour and its microenvironment. Another, more recent study found that preventing the differentiation of prostate myofibroblast cells by using DHA, epithelial to mesenchymal transition (EMT) and tumour invasion (Bianchini et al., 2012). Therefore, interfering with the role of the myofibroblast in the tumour microenvironment should reduce the proliferation and migration of cancer cells.

Studies in breast cancer indicate that the presence of myofibroblasts often correlates with increased invasiveness and poor prognosis. One study by Yazhou et al (2004) showed that loss of CD34 expression and increased expression of α-SMA in myofibroblasts was associated with carcinomas and not with normal breast tissue (Yazhou et al., 2004). Another study observed positive correlations between the presence of myofibroblasts in the tumour stroma and expression of the proliferation marker Ki67 and the proto-oncogene HER-2 in breast cancer cells in all patients tested (Surowiak et al., 2006). In lung adenocarcinoma, high myofibroblast signals were associated with increased lymph node metastasis, high stage tumour growth, vascular invasion and a shorter survival time (Shu and Li, 2012). However there is conflicting evidence among lung cancer studies; a report by Matsubara et al

reported that sub-epithelial myofibroblasts identified by α-SMA expression in lung adenocarcinoma was actually a histological indicator of excellent prognosis in the patients they tested (Matsubara et al., 2009). These varying reports on the role of myofibroblasts in cancer progression demonstrate the continuing need for further research into the role that different forms of myofibroblasts may play in the development of different forms of tumour in different tissues.

As yet the role of the myofibroblasts in gastric cancer is not well documented; studies in scirrhous gastric cancer patients show that an increase in myofibroblast cells is correlated with a worse prognosis (Fuyuhiro et al., 2010c). Studies in mouse models of gastric cancer show that tumour associated myofibroblasts express VEGFA, which contributes to increased angiogenesis (Guo et al., 2008). In addition, a recent paper by Holmberg et al (2012) showed that CAMs increased migration and proliferation of gastric cancer cells, when compared to either ATMs, or ANM. This study also showed that conditioned medium from gastric CAMs was sufficient to stimulate migration, invasion and proliferation of gastric cancer epithelial cells, when compared to media collected from ANMs cells. Also, proteomic analysis of the myofibroblast secretomes revealed a decrease in extracellular matrix adaptor protein like transforming growth factor induced gene-h3 in the cancer myofibroblast. Significantly, this decrease was correlated with lymph node metastasis, worse prognosis and shorter patient survival (Holmberg et al., 2012).

#### 1.4.1.3 Paracrine Communication between CAFs and cancer cells

Many studies have been carried out to elucidate the paracrine signalling that reciprocal changes in CAFs and cancer cells during tumour progression. This form of epithelial-mesenchymal crosstalk drives proliferation of both epithelial and stromal cells (Bhowmick et al., 2004b). Studies into the development of prostate cancer show that cancer progression was detected in genetically initiated epithelial cells but not in non-cancerous epithelial cells, thus indicating that CAFs stimulate progression of cancer but do not initiate cancer development. It is also interesting to note that CAFs changed the histology of the epithelial cells, but did not induce tumorigenesis (Olumi et al., 1999).

### 1.4.2 Cancer stroma models

#### 1.4.2.1 *In vitro/in vivo* models of the cancer microenvironment

To investigate the molecular mechanisms of cross-talk between cancer cells and CAFs, CAFs have been isolated from several different tissues before being studied in isolation, or in co-cultures with cancer cells. These co-culture methods allow specific phenotypes including migration, proliferation, and invasion to be systematically investigated. Some groups have developed self-renewing breast cancer models, consisting of cancer and stromal cells maintained in a constant 1:1 ratio (Piechocki, 2008), while others have developed more elaborate co-culture systems which incorporate extracellular matrix and basement membrane components (Weaver et al., 1995). While these kind of *in vitro* models allow detailed dissection of the molecular mechanisms of crosstalk they are minimal models and may not fully

reflect the complex cocktail of factors that balance paracrine effects in a more complex tissue environment. For this reason, several *in vivo* animal models have been used to study the effect of normal and cancer derived fibroblasts on tumour development. As discussed in section 1.3.3.1, the most aggressive cancers with increased vascular systems were produced in mice when cancer cells were recombined cancer stroma rather than benign stroma (Barclay et al., 2005).

Genetically modified animals have also been used to deduce the *in vivo* effects that individual genes or sets of genes may have on tumour development. In transgenic mice, it has been shown that factors secreted from fibroblasts can directly increase the metastasis of tumours. Transgenic mice lacking the SLOOA4 protein are unable to metastasis, while co-injection of cancer (SL00A4 negative) cells with fibroblasts expressing the SL004A protein led to an increase in tumour metastasis. The SL004A gene is released from CAFs before being taken up by neighbouring cancer cells, where it increases cell motility by interacting with myosin (Grum-Schwensen et al., 2005).

### 1.4.3 Gene expression in stroma

Both *in vitro* and *in vivo* models have been used to provide insight into the molecular processes that operate within the microenvironment of different types of tumour. With the increasing availability of high quality gene arrays microarray analysis has become a common method of studying global gene expression patterns in cancer stromal cell populations. As a result, it is now possible to use a list of only 120 genes to classify breast carcinomas based purely on gene expression patterns observed in associated stromal cells (Symmans et al., 2003). Microarray analysis has

also revealed the range of genes that are differentially expressed in CAFs compared to normal fibroblasts. These differences may provide new insight into the role that re-programmed CAFs play in driving tumour development. Comparison of similar profiles from different tumours or a greater range of patients will reveal the extent to which common mechanisms occur in different forms of cancer or different patients.

Gene expression studies performed on pancreatic cancer and fibroblast cell lines cultured in isolation or as mixed co-cultures show that the COX-2 gene was found to be over-expressed in both cell types and caused increased invasion when cancer cells were co-cultured with CAFs, as a result, COX inhibitors may provide a possible mechanism for therapeutic intervention (Sato et al., 2004). Using the HG-U133A gene chip, fibroblasts derived from the site of a colon tumour or normal regions of the colon were compared. These results revealed that CAFs appeared to be more homogeneous than fibroblasts derived from a histologically normal region. Significantly, this study also showed that fibroblasts were able to stimulate cancer cell proliferation via soluble factors alone, PTGS2 (COX2) and TGF-β, and that the proliferative effect was more intense in fibroblasts derived from tissue closer to the site of the tumour (Nakagawa et al., 2004). In prostate cancer, the MGC-1 gene chip was used to identify 20 genes whose expression profiles correlate with TGF-β1 induced trans-differentiation of fibroblasts into myofibroblasts (Untergasser et al., 2005).

Several groups have used laser dissection methods to isolate samples of cancer cells and associated fibroblasts from a range of tumours, in order to provide insight into

in situ gene expression profiles (Micke et al., 2007). As a result prognostic gene expression signatures have been defined for several types of cancer, including prostate cancer, where 44 genes were found to be consistently differentially expressed in the cancer stroma (Finak et al., 2008). Laser captured micro-dissection methods were also used to compare gene expression profiles in epithelia and stromal cells from patients with either pancreatic cancer or pancreatitis, thus providing a new way to distinguish between the two conditions (Fukushima et al., 2004). In breast cancer, multi-variant analysis methods were used to define a stroma derived prognostic prediction (SDPP) signature. The multi-variant analysis displayed distinct clusters; representing distinct types of cancer. Patients with reduced re-occurrence had hits in a cluster of genes involved in the immune response, while patients with high levels of re-occurrence had hits in a cluster of genes involved in reduced wnt signalling, hypoxia and angiogenesis (Finak et al., 2008).

### 1.4.4 Genetic mutations in stroma

It is well documented that cancer cells continue to acquire somatic mutations during tumour progression. These mutations can confer a selective advantage, which contribute to tumour growth, immune evasion and drug resistance (Komarova and Wodarz, 2005; Sumimoto et al., 2006). The occurrence of mutations in stromal cells is less well understood. A study investigating mutations in epithelial or stroma cells in breast neoplasia found that three genes (TP53, PTEN and WFDC1) showed loss of heterozygosity, in both epithelial and stromal compartments, leading to the suggestion that stromal cells were originally epithelial cells that

underwent epithelial-mesenchymal transition (Kurose et al., 2002). More recent studies suggest that stromal myofibroblasts are genetically more stable than associated cancer cells and do not show significant changes in chromosomal stability or rearrangement, as such CAFs may represent a more attractive therapeutic target as increased genetic stability means that these cells are less likely to evolve resistance to therapeutic drugs (Gururajan et al., 2012).

### 1.4.5 Paracrine signalling in the cancer micro-environment

Tumour development is clearly driven by a process of reciprocal paracrine signalling or cross-talk between cancer and stromal cells (Bhowmick et al., 2004a). Although paracrine signalling is a normal part of tissue homeostasis, it appears that changes in cancer cells modify signals sent to stromal cells. As a result, stromal cells become re-programmed to produce different factors, including a range of growth factors including FGF, IGF, EGF, hepatocyte growth factor (HGF), Platelet derived growth factor (PDGF), vascular endothelial growth factor (VEGF) and TGF-β, all of which, except TGF-β, stimulate epithelial cell proliferation (Bhowmick et al., 2004b; de Jong et al., 1998a).

PDGF is released from many different cell types and its receptor is present on fibroblasts but not epithelial cells. Stimulation of the PDGF receptor (PDGFR) induces cell division, enhanced survival and migration of fibroblast to the site of injury/inflammation (Bonner, 2004). In cervical cancer, the PDGFR inhibitor Imatinib was used to demonstrate the role of PDGF in the cancer microenvironment. PDGF binds to its receptors on fibroblasts and pericytes, causing the expression of FGF-7, which induces epithelial cell proliferation and expression of FGF-2, which then

increases angiogenesis by binding to its receptors on endothelial cells (Pietras et al., 2008).

TGF-β is expressed by epithelial cells, and its receptors (TGF-βR) are expressed by fibroblasts. TGF-β produced by cancer cells was shown to induce trans-differentiation of fibroblasts into myofibroblasts, thereby causing a reciprocal increase in cancer cell growth in co-culture studies. Interestingly the chloride channel CLIC4 was highly up-regulated (x16) in activated fibroblast, CLIC4 inhibits myofibroblast leading to a stationary phenotype (Samoszuk et al., 2005); (Ronnov-Jessen et al., 2002). Equally, expression of a chloride channel in epithelial cells could inhibit breast cancer progression (Gruber and Pauli, 1999). In ovarian cancer, TGF-β induces fibroblasts to secrete fibroblast activation protein (FAP), which in turn increases cancer cell invasion, proliferation and migration (Chen et al., 2009). Similar findings in prostate cancer show that TGF-β increases myofibroblast numbers, angiogenesis and tumour growth (de Jong et al., 1998b; Tuxhorn et al., 2002b). TGF-β is complex and has dual opposing roles, with dual time dependent autocrine effects on cancer cells. Firstly, inhibiting the DNA synthesis effects of EGF, FGF and PDGF produced by cancer cells. Then, after a 24 hours, TGFβ itself stimulates DNA synthesis within cancer cells (Shipley et al., 1985). TGF-β also has opposing roles in cancer progression. In addition to increasing fibroblast proliferation, migration and invasion of cancer cells, TGF-β also demonstrates inhibitory effects on the growth and invasive properties of cancer cells. Knocking out the TGF-β receptor in fibroblasts resulted in a two-fold increase in tumour growth and invasion. TGF-β inactivation leads to an increase in expression of

tumour growth factor alpha (TGF-α), macrophage stimulating protein (MSP) and HGF from fibroblasts, which activate the ERB, RON and c-met receptors on epithelial cells, respectively (Cheng et al., 2005).

HGF expressed by fibroblasts acts to weaken cell-cell contacts and induce the degradation of the extracellular matrix thereby allowing increased migration of cancer cells. In low oxygen conditions, epithelial cells increase expression the HGF receptor (c-met), thereby increasing the sensitivity of cancer cells to HGF within the tumour microenvironment. Stimulation of the c-met receptor induces a distinct change in epithelial cell morphology, displaying a fibroblastic cell shape, which is characteristic of epithelial-mesenchymal transition and increased cell scattering. HGF expressed from fibroblasts is also known to act as an angiogenic factor, which acts through the c-met receptor on endothelial cells. However, HGF also promotes angiogenesis independently of the met receptor by causing blood vessels to branch (Michieli et al., 2004). Increased levels of HGF and HGF-receptor have been observed in the late stages of several tumours and this signature is clearly linked to 'bad' patient prognosis (Aune et al., 2011). In addition, in mice truncation of HGF protects cells against radiotherapy, while truncation of the met receptor increased sensitivity to radiotherapy (Michieli et al., 2004). In gastric cancer, PGE2 and wnt signalling pathways are activated in epithelial cells, causing the release of an unknown soluble factor, which activates fibroblasts and recruitment of bone marrow derived mesenchymal stromal cells (MSCs), which differentiate into myofibroblasts. Activated myofibroblasts then secrete high levels of VEGF, which increases angiogenesis (Guo et al., 2008).

### 1.4.6 Extracellular matrix

The secretion of extracellular matrix proteins and extracellular proteases from fibroblasts aids tumour cell invasion. In breast cancer, media collected from CAFs treated with TGF-β, significantly increased the invasiveness of cancer cells by causing fibroblasts to secrete extracellular matrix proteins fibronectin and laminin (Casey et al., 2008). In general, extracellular matrix proteases allow degradation of the basement membrane to allow epithelial cells access to the stroma, whilst the secretion of extracellular matrix proteins allows fast migration of epithelial cancer cells along 'fibre roads', aiding travel through the stroma (Casey et al., 2009).

Tight regulation of the secretion of proteases and protease inhibitors, is essential for normal physiological processes, however, during tumour development this balance is disrupted (Noel et al., 2008). Matrix metalloproteinases (MMPs) are expressed by fibroblasts, endothelial cells, macrophage-like cells and cancer cells. Tumour cells induce expression and secretion of MMPs, which degrade all extracellular matrix proteins facilitating tumour cell invasion. The cocktail of MMPs produced in the tumour microenvironment includes collagenases, which degrade fibrillar collagen; gelatinases, which degrade denatured colllagen and proteglycans or glycoprotein, such as laminin, fibronectin and gelatine (Heppner et al., 1996). MMPs facilitate EMT, by degrading connections between epithelial cells such as E-cadherin, therefore facilitating epithelial movement (Noe et al., 2001). MMP3 in particular has been shown to play a key role in EMT (Sternlicht et al., 1999).

Fibroblasts express all MMP family members apart from MMP7, which is uniquely expressed by epithelial cells (McCaig et al., 2006a). *H. pylori* infection increases

MMP7 expression causing an increase in the proliferation and migration of stromal myofibroblasts, through activation of PI3K signalling cascades (Hemers et al., 2005). In addition, insulin-like growth factor two (IGF-II) and insulin-like growth factor binding protein five (IGFBP5) are both expressed by fibroblasts. MMP7 cleaves IGFBP5 causing the release of IGF-II, which itself stimulates proliferation and migration of epithelial cells (McCaig et al., 2006b). Finally, MMP-9 is released from fibroblasts in many cancers in response to TGF-β stimulation (Stuelten et al., 2005).

Urokinase – type plasminogen activator (uPA) is a serine protease expressed by myofibroblasts and has a strong associations with cancer progression (Nielsen et al., 1996; Smith and Marshall, 2010). uPA is activated upon binding to its receptor (uPAR) on epithelial cell, fibroblasts and macrophages (Dublin et al., 2000a). Once bound to its receptor, uPA converts inactive plasminogen into plasmin, a serine protease that degrades fibrin and activates MMPs (Carmeliet et al., 1997), leading to the degradation of the extracellular matrix and increased invasion and metastasis of cancer cells. Furthermore, expression of plasminogen activator inhibitors 1 and 2 is decreased in many cancers, including gastric (Nakagawa et al., 2004; Norsett et al., 2010). As such, high levels of uPA and its receptor are indicative of poor prognosis in gastric cancer (Heiss et al., 1995).

### 1.4.7 Therapeutic interventions

In most cases, gastric cancer is only detected at late stages of tumour development -due to the lack of early stage symptoms. A combination of surgery, chemotherapy and immunotherapy are most commonly used to treat the tumour, however 5-year survival figures remain poor, with most combination therapies only increasing life

expectancy of late stage gastric cancer patients by around six months (Vecchione et al., 2009). As such, there is a real need for less invasive stage-specific, or even patient specific, cancer drugs. Given the importance of epithelial mesenchymal cross talk in tumour progression, the microenvironment has become a focus for the development of new therapeutic strategies.

Due to the tumour promoting roles of PDGF, it seemed an attractive target for therapy, but as there are organ specific forms of PDGF receptor, development of specific drugs has proved problematic (Bonner, 2004). The drug Glivec (also known as Imatinib or Gleevec) was the first tyrosine kinase receptor inhibitor approved as an anti-cancer drug, it has been developed to inhibit PDGF-$\alpha$ and $\beta$ receptors. Glivec works in two ways, inhibiting the PDGFR on fibroblasts inhibits PDGFs mitogenic properties and inhibiting PDGFR on endothelial cells inhibits PDGF stimulated VEGF release. Glivec has also been shown to lower high intestinal fluid pressure (often associated with cancers) therefore increasing the amount of cancer drugs absorbed (Capdeville et al., 2002). Glivec also inhibits the effects of PDGF on fibroblasts, inhibiting the release of FGF-2 and FGF-7, therefore decreasing vascularisation and epithelial proliferation respectively (Pietras et al., 2008). Glivec is currently approved in several countries, for treatment of post-surgery gastro-intestinal stromal tumours.

Marimastat is an inhibitor of MMPs, which was trialled as a possible anti-cancer drug in several different organs. However, in each case mariamastat failed to have beneficial effects, with the exception of gastric cancer (Fielding.J, 2000). Drugs targeting the uPA receptor have also proved to be unsuccessful; this is thought to

be due to off-target effects, such as increased cell invasion, proliferation and survival. Combined treatment with truncated of HGF and a truncated HGF receptor has been suggested for future cancer therapy. In addition, combining the truncated met receptor with radiotherapy may be another possible therapeutic strategy, as in mice this increased cancer cells sensitivity to radiotherapy (Michieli et al., 2004). New strategies that interfere with paracrine signalling or stroma/cancer metabolism may offer new hope for more effective intervention even in late stage tumours. However, in order to develop more rationally designed therapeutic strategies we need to develop a better understanding of the molecular processes and pathways that regulate paracrine effects in both cancer cells and associated stromal cells. Given the complexity of the systems involved, it is unlikely that answers to these questions will be provided by conventional reductionist approaches that focus on the mechanistic properties of individual cellular components.

## 1.5 Systems biology

Growing realisation of the need to understand the broader properties of biological systems has led to a shift away from classical reductionist methods towards more global or 'systems level' approaches, leading to a new field of network biology, and more recently network medicine. Unlike classical textbook representations, biological processes are inherently non-linear dynamic processes often involving multiple combinatorial or conditional steps or components. In this context, it is highly unlikely that one gene could be responsible for a specific phenotype. Even in situations such as cystic fibrosis where this is true, it is the combination of global gene expression profiles that determine the severity and rate of progression in

different patients. Furthermore, even if defects are observed in a specific pathway component, this does not mean that the same component is a realistic therapeutic target. Better options may exist either up or downstream of the defective component however, rational selection of better therapeutic candidates requires insight into global network structure and an ability to predict potential effects on other important processes.

### 1.5.1 Protein interaction networks

Mathematicians have been applying graph theory since 1736, in this approach objects are plotted as 'nodes' and their relationships or interactions with one another as 'edges'. Use of graph theory to map biological functions, allows visualisation of the interconnectivity between different biological processes. Over the past decade graph theory has been applied to systematically map out all biological functions and evaluate topological features (Barabasi and Oltvai, 2004). In biological networks, nodes normally represent proteins, metabolites or enzymes, whilst edges between them represent interactions or chemical reactions. Biological networks can be un-directed, with edges being represented by lines, or directed, with edges being represented by arrows (Guimera and Nunes Amaral, 2005).

Many different topological measures are used throughout Network Biology to understand the control that a node, or a cluster of nodes, have over other network members. The degree of a node represents the number of interacting partners within the network, therefore nodes with a higher degrees have a greater ability to effect more members and/or a greater number of processes.

#### 1.5.1.1 Topological features

Unlike random mathematical networks, biological interaction networks tend to have a scale-free topology, with a few highly connected central proteins, and many less connected nodes. Removal of the less connected nodes has relatively little effect on network function, whereas highly connected nodes, called 'hubs,' tend to be essential. Removal or mutation of these nodes in yeast frequently led to network fragmentation, and lethality. In general, a scale-free topology offers protection against random mutations, which would most likely affect the least connected genes (Jeong et al., 2001; Li et al., 2008). In general, biological networks are not homogeneous, they contain areas of dense connectivity surround by areas with fewer connected nodes. These dense areas are termed modules. However, the biological significance of modules within protein interaction networks has been disputed; some believe that modules are just artefacts of network coverage, whilst others suggest modules are functional and represent an important level of cellular organisation.

In cancer research, modules within protein-protein interaction networks correctly identified GO biological function (Chuang et al., 2007). A functional module is a subset of molecules that are more connected or related to one another than other molecules due to their molecular/chemical specificity, cellular localisation or expression profile. This modular structure, allows changes within a module to occur without affecting the entire network. Similarly changes in connections between modules may affect multiple different processes. In this sense the modular nature of biological networks may facilitate functional evolution, which could not occur if

molecules where un-structured, or non-modular (Hartwell et al., 1999). Many software tools are now available to analyse the topological feature of networks. However, although each algorithm identifies clusters of proteins/genes rational evaluation is often required to evaluate the biological significance of the identified clusters (Hallinan, 2004). Netbox and MCODE both define modules based on network topology, Netbox is an online resource that allows users to map gene expression data onto a human interaction network in order to identify unique linker genes, network modules and perform comparisons with genes identified in previous cancer studies (Bader and Hogue, 2003; Cerami et al., 2010). This computational tool was initially developed to identify driver genes from passenger genes (Greenman et al., 2007) as identification of similar modules across a range of patients with a particular disease, may indicate new 'driver genes'. Interestingly, in different types of cancer, although overall the same modules were affected, different members were changed in different cancers. Therefore, knowledge of altered cancer related modules may help develop better intervention strategies (Cerami et al., 2010).

The importance of network biology in understanding cancer was highlighted by a study that used network-based analysis to identify genes involved in breast cancer metastasis. Gene expression arrays were mapped onto protein-protein interaction networks, producing sub-networks common to metastasis. One particular gene was not significantly changed but was found to be essential for inter-connecting a specific sub-graph, which was important in predicting whether a sample would be metastatic. The un-changed gene is similar to the linker genes identified in Netbox.

Gene array analysis alone would not have identified the importance of these genes, but the combination with protein-protein interaction data highlights the future importance of this form of network biology in cancer research (Chuang et al., 2007).

A number of studies have highlighted the importance of highly connected 'hub' proteins and their positioning either within or outside of network modules. In the yeast interactome, two types of hubs were observed, party hubs and date hubs. Party hubs displayed a similar pattern of co-expression with their interacting partners, and tended to share the same sub cellular localisations as their partners. In contrast, date hubs tend to exhibit lower levels of co-expression or co-distribution with their interacting partners. As such, party hubs tend to occur within biological modules, while date hubs are important to the overall structure of the network and tend to occur between modules. Removal of party hubs did not affect the network connectivity, whilst the removal of date hubs resulted in smaller sub-networks each with specific biological functions (Han et al., 2004).

Another study also assumes that the position of a node within a modular network may determine its role. In this study, nodes within modules were first categorised as a hub or non-hub, then using a participation co-efficient to determine how well-connected nodes are to other modules, hubs were divided into 3 categories and non-hubs into 4 categories. Results from this analysis suggest that although most modules are associated with one predominant function, several have multiple functions, indicating over-lapping central metabolic pathways. As predicted hubs have the lowest evolutionary loss rate compared to more peripheral nodes. The surprising finding is that non-hub connectors, genes that do not have a high degree

but are found between modules are significantly more conserved than hubs that are highly connected within a module. This may be because the removal of a hub within a module may be backed-up by redundant interactions, whilst the removal of a non-hub connector may be catastrophic to the communication through module-module connections in the network. Therefore, this study supports previous findings that show hubs to be more essential in the network, but also suggests that there may be other important non-hubs in the network which are better identified by global connectivity algorithms rather than simple degree (Guimera and Nunes Amaral, 2005).

To connect two nodes on a network, there may be a number of possible paths passing through different nodes, each being of variable lengths. The shortest path is defined as the shortest route that connects two nodes, and this is often used to calculate betweenness centrality values for nodes within a particular network (see below). Path-length is a quantitative measurement, and is calculated as the number of nodes that lie on the shortest path between two nodes. The clustering co-efficient represents the tendency of a node to form groups, as shown in Figure 1.3, if A and B interact, and B and C interact, A and C may interact if they have a high clustering co-efficient. The clustering co-efficient of A is higher than the clustering co-efficient of node F, as 2 of node A's interaction partners interact, whilst none of node F's partners interact (Barabasi and Oltvai, 2004).

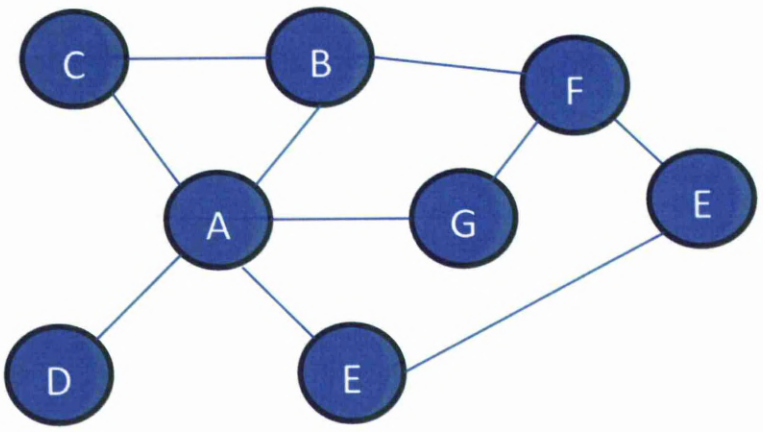



**Figure 1.3 Network representing the increased clustering co-efficient tendency of certain nodes.**

Betweenness centrality is a measure of the amount of control that a node may have within a network. The concept of node centrality was first proposed by Bavelas et al in 1948 and is calculated by taking into account the number of shortest paths that pass through a node to connect two other nodes. A node that lies on a large number of shortest paths, is considered central and is thought to have more control over network function by inhibiting or increasing information flow to other parts of the network (Bavelas, 1948) (Freeman, 1977), with high betweenness nodes are known as 'bottleneck' nodes (Missiuro et al., 2009).

### 1.5.2 Disease networks

Human protein-protein interaction networks have been used to understand the complex biological interactions, occurring in different diseases. This understanding of networks is useful in drug-discovery, as understanding the entire network of possible circuits, allows the researcher to predict the consequences that targeting a single gene may have, including possible off-target effects or neutralising circuits (Zanzoni et al., 2009). In addition, the concept of pleiotrophy in disease is

important; a single gene that has functions in different pathways can have multiple 'pleiotropic' effects. In this context differences in function result from a single protein existing in different functional states or being part of multiple complexes involved in different cellular processes. Genes encoding this class of protein are thought to be genetically more stable than genes encoding proteins with roles in a single process, because a greater selection pressure will be imposed on a gene encoding proteins required for multiple processes. In this context, network biology has been useful in understanding the different disease phenotypes observed upon mutation of one gene. Different in-frame mutations do not always cause the 'removal' of gene activity but may cause a loss of binding or activity with another specific gene or protein. Different in-frame mutations can represent the removal of edges from a network, rather than a node, therefore its activity is altered rather than removed (Zhong et al., 2009).

A set of human disease networks were produced to help identify groups of similar diseases based on their individual disease genes, at a network level. The networks consist of either, diseases that are connected to each other if they share common mutated genes, or genes that are connected to each other if they are related to similar diseases. Genes within distinct disease modules represented functional GO processes and interacted with each other more than random chance, which suggested the use of the network as a possible 'diseasome' for future studies. Interestingly in these studies hubs were shown to weakly correlate with disease genes. This is most likely due to the fact that disease genes can be split into essential and non-essential genes. Hubs are often essential genes (Jeong et al.,

2001) and the majority of disease genes are shown to be non-essential. The reason the majority of germ line disease genes are non-essential could be explained using an evolutionary theory; if a disease gene was central it may lead to lethality, while peripheral disease genes are less likely to have such catastrophic effects and will be passed on through the generations (Goh et al., 2007).

### 1.5.3 Cancer networks

A complex relationship between genetics, epigenetics and environmental factors are thought to contribute to the progression of cancer. As outlined above, a systems level approach is required to deal with the inherent complexity of cancer networks. Currently available cancer drugs that target a specific receptors or growth factors have shown only small levels of success. In order to improve the success of chemotherapy we need to use new systems level approach, rather than the 'target and kill' strategy that has been applied to date. Also, systems based methods need to be used to investigate cancer in a stage specific manner, as cells and their environment operate in 'dynamic equilibrium' which changes throughout cancer progression (Knox, 2010).

Network of cancer genes is an on-line web resource, containing a collection of 736 human genes that are mutated in some forms of cancer(Syed et al., 2010). Around half of these genes come from the Cancer Gene Census and the remaining set from high-throughput mutation cancer screenings. The resource supports a systems based approach, providing protein interaction information for each cancer related gene independent of their actual functions. 579 of the 736 cancer genes can be mapped onto the protein-protein interaction network, from which topological

calculations of degree, clustering co-efficient and betweenness have been calculated.

#### 1.5.3.1 Topological features of cancer networks

Many different studies have been carried out to identify topological features of proteins within cancer related networks, with the aim of identifying mutations that upon may have the biggest effect on tumour progression or prognosis.

As the scale-free topology of biological networks should protect against genetic mutation, alterations in epigenetic regulation could play an important role in cancer phenotypes. Interestingly, epigenetic alterations have been shown to have stronger association with cancer than genetic mutations. Also, diets that stabilise epigenetic proteins have been shown to have a positive protective effect over cancer progression (Knox, 2010). It appears that in general cancer related proteins are significantly more connected within interaction networks than would be expected by chance, with most having a high degrees in the network, suggesting that cancer proteins play a central role in the network function (Jonsson and Bates, 2006). However, many people worry that protein-protein interaction networks are bias towards cancer studies, and that the observed increase in connectivity involving cancer related proteins, may reflect study bias. In defence, Jonsson *et al* show cancer related proteins tend to contain significantly more promiscuous interaction domains in relation to non-cancerous proteins, with domain data retrieved from PFAM showing no bias towards cancer proteins (Jonsson and Bates, 2006).

#### 1.5.3.1.1 Cancer mutated genes and hubs

Some cancer related mutations clearly occur in germline cells such as p53 mutations, as such these are conserved inheritable mutations (Jonsson and Bates, 2006). However, many more cancer related mutations are somatic, occurring in differentiated cells and these are rarely inherited (Zanzoni et al., 2009).

The fact that most cancer genes are not conserved through evolution would explain the high number of cancer related hub proteins. While 5% of nodes in a random network would be expected to be hub proteins, over 13% of hubs in the current human interactome are cancer genes, therefore demonstrating an enrichment of hubs in cancer related genes (Syed et al., 2010). In addition, cancer related genes tend to be associated with signalling hubs (Awan et al., 2007; Cui et al., 2007). This work has led to the establishment of an oncogenic signalling map and identification of 'oncogenic super-highways, which are often mutated in cancer.

### 1.5.3.2 Crosstalk in biological networks

Superimposing canonical pathways onto protein-protein interaction networks demonstrates a large amount of potential cross-talk between pathways (Wu et al., 2010). One particular study used protein-protein interaction data to investigate crosstalk between different significantly changed biological pathways. They produced a pathway crosstalk network in which clusters of pathways were shown to share similar biological functions, with signalling related pathways generally producing the largest cluster. They concentrated on identifying pathways that had low numbers of cross-talk between them, in order to identify cross-talk between

less similar processes and state the networks could be used as a framework for datasets to be mapped onto (Li et al., 2008).

Currently, tools such as NetBox are available which allow the identification of modules within networks, a prediction is then made of the range of possible biological pathways that individual modules could be involved in. Alternative tools, such as Metacore™, allow the identification of over-represented pathways within given datasets but do not provide information on connectivity between pathways. There is currently no tool available that allows the user to map differentially regulated genes on to pathways, identify over-represented pathways, and identify the entire range of pathway crosstalk within an interaction network.

#### 1.5.3.2.1 Personalised medicine

In the future, individual patient genome sequence information will be more affordable, by combining this information with interactome and expression data, it will be possible to use network biology methods to provide new insights into personalised or stratified medicine and optimal strategies for cancer treatment (Zanzoni et al., 2009). In the meantime, patient prognosis scores have been used to try and predict the stage and type of cancer, in order to administer the appropriate treatment. In gastric cancer, one group devised a method for scoring patients into prognostic groups. cDNA samples from cancer and normal regions were screened for five conventional pathological factors; tumour size, depth of invasion, histological growth pattern, lymph node metastasis and liver metastasis. Applying a t-test to each pathological factor, they show that patients could be divided into 'good' and 'bad' prognosis groups, with the scoring system successfully

differentiating, previously hard to predict patients, which are at high risk (Inoue et al., 2002). Similar pathological factors were utilised by an alternative group, who defined a prognostic scores for gastric cancer (PSGC) using a ten-variable staging multivariate technique (Kologlu et al., 2000). Our patient samples have associated prognosis scores assigned based on a similar list of staging variables, which we aim to devise an appropriate method to deduce outcome. To conclude, although there is a large number of cancer interaction networks currently published, the important relationship between the cancer and stromal cells is evident, and the current requirement for improved stromal/cancer networks is clear. In addition, the combination of patient prognosis data, used together with techniques enabling differentially regulated genes to be mapped to over-represented pathways, within cross talk visualising, protein-protein interaction networks, may provide us with the best approach to be able to understand the complex processes occurring within the cancer micro-environment, within different tumour stages.

The first two results chapters within this thesis apply a range of canonical pathway and statistical tools to compare the gene expression profiles between cancer-associated, adjacent and absolute normal myofibroblasts. Key molecular signatures and biological processes were identified, such as the differential regulation of genes involved in cell adhesion, glycan metabolism, DNA repair and metabolic pathways including fatty acid β-oxidation, ketone body synthesis and cholesterol biosynthesis. Upon receiving additional patient prognosis information, and access to additional gene expression analysis software, several patients were clear outliers and were consequently removed. Chapter five consists of analysis of the refined dataset,

canonical pathway analysis and refined multivariate techniques which confirm the molecular signatures and biological processes identified within the first two chapters, and are expanded to define a system whereby cancer cells undergo metabolic reprogramming to induce a variation of the reverse Warburg effect. Finally, utilisation of the patient prognosis scores are utilised to define prognosis specific expression profiles, thereby providing new insight into the molecular processes that drive important paracrine communication networks during the development of gastric tumours.

# 2 Chapter Two: Methods

## 2.1 Data processing

Biopsies were dissected from patients with gastric cancer in Szeged Hospital, Hungary. Samples were taken from 14 patients with gastric cancer; samples taken directly from the region of the tumour are referred to as cancer samples or Cancer Associated Myofibroblasts (CAMs). In 12 of the 14 cancer patients, biopsies were also taken from tissue adjacent to the site of the tumour, these are known as Adjacent Tumour Myofibroblasts (ATMs). For comparison, absolute normal reference samples were obtained from 12 post-mortem organ donors who had no known underlying medical conditions, which are referred to Absolute normal samples (ANs). All samples are shown in Table 2.1.

Gastric cancer patient follow-up scoring details and related patient details are provided in Table 2.2 and Table 2.3 respectively. Patient prognosis scores are determined using the semi-quantitative scoring system. This scoring system takes into account degree of anaemia, BMI loss, the stage of the tumour (grade, type, lymphatic vessel invasion, vascular invasion and positive margins of resection), tumour marker elevation (serological), re-occurrence of the tumour, whether the tumour was a synchronous or metachronous, metastasis and the mortality of the patients. Based on these measures histologists in Hungry then assigned patient prognosis scores (Table 2.4) from 0-14, with scores of 0 relating to no difference from a healthy subject and 14 representing a very progressive stage of gastric cancer. Patient sample details associated with absolute normal patients are provided within Table 2.5.

| Label | Sample | Sample type |
|---|---|---|
| 1-CAM | Sz42/1 P5 | Cancer |
| 1-ATM | Sz42/2 P5 | Adjacent |
| 2-CAM | Sz45/1 P5 | Cancer |
| 2-ATMA | Sz45/2 P5 | Adjacent 1 |
| 2-ATMB | Sz45/22 P7 | Adjacent 2 |
| 3-CAM | Sz190/1 P4 | Cancer |
| 3-ATM | Sz190/2 P4 | Adjacent |
| 4-CAM | Sz192/1 P5 | Cancer |
| 4-ATM | Sz192/2 P5 | Adjacent |
| 5-CAM | Sz194/1 P5 | Cancer |
| 5-ATM | Sz194/2 P5 | Adjacent |
| 6-CAM | Sz195/1 P5 | Cancer |
| 6-ATM | Sz195/2 P5 | Adjacent |
| 7-CAM | Sz198/1 P5 | Cancer |
| 7-ATM | Sz198/2 P5 | Adjacent |
| 8-CAM | Sz268/1 P5 | Cancer |
| 8-ATMA | Sz268/2 P5 | Adjacent 1 |
| 8-ATMB | Sz268/22 P5 | Adjacent 2 |
| 9-CAM | Sz271/1 P5 | Cancer |
| 9-ATM | Sz271/2 P5 | Adjacent |
| 10-CAM | Sz294/1 P4 | Cancer |
| 10-ATMA | Sz294/2 P5 | Adjacent 1 |
| 10-ATMB | Sz294/22 P4 | Adjacent 2 |
| 11-CAM | Sz305/1 P5 | Cancer |
| 11-ATMB | Sz305/22 P5 | Adjacent 1 |
| 11-ATMA | Sz305/2 P5 | Adjacent 2 |
| 12-CAM | Sz308/1 P5 | Cancer |
| 12-ATM | Sz308/22 P6 | Adjacent |
| 13-CAM | Sz187/1 P8 | Cancer |
| 14-CAM | Sz197/1 P5 | Cancer |
| 15-CAM | Sz389/1 P7 | Cancer |
| 15-ATM | Sz389/2 P7 | Adjacent |
| 21-ANMA | Sz196/2 P5 | Absolute normal |
| 22-ANMA | Sz241/2 P6 | Absolute normal |
| 22-ANMB | Sz241/22 P6 | Absolute normal |
| 23-ANMA | Sz246/2 P6 | Absolute normal |
| 23-ANMB | Sz246/22 P6 | Absolute normal |
| 24-ANMA | Sz261/2 P6 | Absolute normal |
| 24-ANMB | Sz261/22 P6 | Absolute normal |
| 25-ANMA | Sz279/22 P4 | Absolute normal |
| 26-ANMA | Sz334/2 P5 | Absolute normal |
| 26-ANMB | Sz334/22 P5 | Absolute normal |
| 27-ANMA | Sz351/2 P5 | Absolute normal |
| 27-ANMB | Sz351/22 P5 | Absolute normal |
| 28-ANMB | 845P7 | Absolute normal |

**Table 2.1.** All cancer, adjacent and absolute normal samples. Table shows the labels used throughout analysis, which correspond to a patient sample ID, and the type of sample. Patient samples shown in red were removed, but are shown to explain the gaps in patient numbering.

| | | Score | | | Score |
|---|---|---|---|---|---|
| Anaemia | HGB<100g/L or blood transfusion | 0 | Histological type | Intestinal | 0 |
| | HGB>100g/L | 1 | | Diffuse | 1 |
| BMI | BMI<19 or Bodyweight loss >10kg | 1 | Lymphatic vessel invasion | Yes | 1 |
| | BMI>19 | 0 | | No | 0 |
| TNM staging classification | T1 | 1 | Vascular Invasion | Yes | 1 |
| | T2 | 2 | | No | 0 |
| | T3 | 3 | Positive margins of resection | Positive | 1 |
| | T4 | 4 | | Negative | 0 |
| | N0 | 0 | Tumour markers elevation | Yes | 1 |
| | N1 | 1 | | No | 0 |
| | N2 | 2 | Tumour reoccurrence | No | 0 |
| | N3 | 3 | | With surgical intervention | 1 |
| | M0 | 0 | | Without surgical intervention | 1 |
| | M1 | 1 | Synchronous or metachronous tumour | Yes | 1 |
| Grade | Grade 1 | 1 | | No | 0 |
| | Grade 2 | 2 | Mortality (exit) | Yes | 1 |
| | Grade 3 | 3 | | No | 0 |

**Table 2.2.** A semi-quantitative scoring system was established according to the following parameters, anaemia, loss of bodyweight, TNM (T=size of original tumour, N=nearby lymph nodes, M= distant metastasis), stage of tumour, histological results (grade, histological type, lymphatic vessel or vascular invasions, positive margins of resection), results of the oncological and surgical follow-up (recurrence of tumour –serological tumour marker –elevation, radiological signs of recurrence or re-interventions) and mortality (exit). Scores of 0 represent no detectable difference from healthy subjects, with increasing numbers representing increasing severity.

| Sz187 | Sz197 | Sz389 | Sz192 | Sz45 | Sz195 | Sz294 | Sz268 | Sz271 | Sz194 | Sz308 | Sz305 | Sz190 | Sz198 | Sz42 | Patient code | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CagA +ve | CagA +v | None | CagA +ve | None | CagA +ve | None | CagA +ve | None | *H.pylori* | *H.pylori* | CagA +ve | None | *H.pylori* | None | *H.pylori* status. H.pylori = infected CagA +ve =cagA strain detected None = No H.pylori infection | |
| F | M | M | F | M | F | F | M | M | M | M | M | F | M | M | Gender | |
| 39 | 54 | 67 | 49 | 82 | 85 | 84 | 76 | 72 | 76 | 51 | 59 | 65 | 77 | 72 | Age | |
| TOTAL | TOTAL | DISTAL | TOTAL | DISTAL | DISTAL | TOTAL | TOTAL | TOTAL | DISTAL | TOTAL | TOTAL | TOTAL | DISTAL | TOTAL | Type of resection | |
| 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | Anaemia HGB<100G/L or Blood Transfusion | |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | BMI <19 or Bodyweight loss >10KG | |
| 4 | 2 | 3 | 3 | 3 | 1 | 3 | 4 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | pTNM | T |
| 2 | 0 | 1 | 2 | 2 | 1 | 0 | 2 | 1 | 0 | 2 | 2 | 0 | 0 | 0 | | N |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | | M |
| 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 1 | 3 | 3 | 3 | 2 | 3 | Grade | |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | Histological type: Intestinal or diffuse | |
| 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | Lymphatic vessel invasion | |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | Vascular invasion | |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | Positive margins of resection | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | Serological tumour marker-elevation | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | Tumour recurrence | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | Synchronous or Metachronous tumour | With surgical intervention |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | | Without surgical intervention |
| 15.5.07 | 16.8.07 | 2.7.06 | 21.5.07 | 26.6.06 | 7.8.07 | 5.12.07 | 6.11.07 | 13.11.07 | 6.8.07 | 8.1.08 | 3.1.08 | 18.5.07 | 27.8.07 | 16.6.06 | Date of operative | |
| 42 | 62 | 50 | 22 | 2 | 13 | 51 | 15 | 24 | 60 | 9 | 17 | 5 | 60 | 75 | Survival (months) | |
| | | | 29.3.09 | 8.06 | 14.9.07 | 12.3.12 | 2.09 | | 1.9.12 | 19.10.08 | 1.8.09 | 14.10.07 | 10.9.12 | | Date of Death | |
| 11 | 6 | 7 | 11 | 11 | 9 | 7 | 11 | 12 | 4 | 12 | 12 | 13 | 7 | 5 | Total | |
| 6 | 6 | 6 | 4 | 6 | 6 | 6 | 6 | 5 | 6 | 3 | 6 | 6 | 4 | 6 | Myoscore | |

**Table 2.3.** Patient details relating to the scoring details described in Table 2.2. Row displaying the total represents the progonosis score and is calculated based on the sum of all the variables. If the value is left blank the patient is not deceased.

| Label | Sample | Age | Gender | Location of Tumor | Lauren Classification | Prognostic Score |
|---|---|---|---|---|---|---|
| 1 | sz42 | 72 | M | antrum corpus border | medullar (non-Lauren) | 5 |
| 2 | sz45 | 82 | M | antrum | intestinal | 11 |
| 3 | sz190 | 66 | F | antrum corpus border | mixed | 13 |
| 4 | sz192 | 50 | F | antrum corpus border | diffuse | 11 |
| 5 | sz194 | 76 | M | antrum | intestinal | 4 |
| 7 | sz198 | 77 | M | antrum | intestinal | 7 |
| 8 | sz268 | 76 | M | antrum | intestinal | 11 |
| 9 | sz271 | 72 | M | corpus | mixed | 12 |
| 10 | sz294 | 84 | F | antrum corpus border | intestinal | 7 |
| 11 | sz305 | 59 | F | antrum and corpus | diffuse | 12 |
| 12 | sz308 | 51 | M | antrum corpus border | mixed | 12 |
| 15 | sz389 | 67 | M | antrum | intestinal | 8 |
| 13 | sz187 | 39 | F | antrum corpus border | intestinal | 11 |
| 14 | sz197 | 54 | M | antrum corpus border | diffuse | 6 |

**Table 2.4** Patient sample information. Higher scores in right hand column represent poorer patient prognosis.

| Patient No. | Origin | Age | Gender |
|---|---|---|---|
| Sz196/2 | gastric corpus | 67 | M |
| Sz279/22 | gastric antrum | 60 | M |
| Sz334/2 | gastric corpus | 52 | F |
| Sz334/22 | gastric antrum | 52 | F |
| Sz351/2 | gastric corpus | 41 | M |
| Sz351/22 | gastric, antrum | 41 | M |
| Sz241/2 | gastric, corpus | 44 | F |
| Sz241/22 | gastric, antrum | 44 | F |
| Sz246/2 | gastric, corpus | 45 | M |
| Sz246/22 | gastric antrum | 45 | M |
| Sz261/2 | gastric, corpus | 52 | F |
| Sz261/22 | gastric antrum | 52 | F |

**Table 2.5.** Absolute normal (AN) patient details obtained from 12 post-mortem organ donors who had no known underlying medical conditions.

### 2.1.1 Myofibroblast cell culture generation

Myofibroblasts were isolated from tissue samples and cultured by Peter Hegyi, (Department of Medicine, University of Szeged, Hungary), as previously described (McCaig et al., 2006b). Primary myofibroblasts were cultured by Dr I. Steele and Dr C. Holmberg, (University of Liverpool, UK) in a Dulbecco's modified eagles medium (DMEM) (Sigma, Poole, Dorset, UK) supplemented with 10% Fetal Bovine Serum (FBS) (Perbio, Cheshire, UK), 2% Antibiotic-Antimytotic (Sigma, Poole, Dorset, UK), 1% Penicillin-Streptomycin Solution (Sigma, Poole, Dorset, UK), and 1% non-essential amino acid solution (Sigma, Poole, Dorset, UK). Cells were maintained at 37˚C and 5% $CO_2$ and media was changed every 46-60 hours. Confluent cells were split by adding 0.25% typsin (Sigma, Poole, Dorset, UK) for 10-15 minutes. Cells passages were recorded and only cells from passages 4-12 were used for further analysis, as generally after passage 12, their morphology changes and they become senescent.

### 2.1.2 Gene expression array and normalisation

Cells were cultured and RNA extracted using the RNeasy Kit (Qiagen) by Dr I. Steele. Gene expression analysis was conducted on all samples using the GeneChip® Human Genome U133 Plus 2.0 Array (Affymetrix, SantaClara, CA, USA), at the Liverpool Genome Research Facility by Dr Lucille Rainbow. The Human Genome U133 plus 2.0 Array, allows analyses 38,700 gene transcripts 11 pairs of probes are used to measure the level of expression of a single gene. These consist of perfect and mis-match probes, plus a selection of normalisation probes, which consist of a range of human maintenance genes, which have constant levels of expression and are therefore used to normalise signals across multiple arrays. The GeneChip® Scanner 3000 was used to image arrays, and quality control was performed using Affymetrix microarray 5 QC metrics. Gene-arrays were analysed for their quality using the Array Quality Metrics package, which is explained in detail within section 2.5.1.1. Statistical analysis of gene expression profiles was performed in GeneSpring GX.10 (Agilent) with baseline transformation and experiments were normalised using the Mas5 method. This process also flags the data with P (present), M (marginal) and A (absent) flags. For an oligonucleotide to be classed as detected in this study it had to have been assigned a 'present' flag for all patients within one of two comparison groups. This gene list was then analysed for differential regulation as explained in section 2.1.3.2 below.

However, to understand the effect a particular normalisation technique may have on the results, gene expression profiles were also normalised using the increasing popular Robust Multi-array Average (RMA) technique. The RMA derived differentially regulated gene lists were not used within any downstream analysis,

yet we believe it is important to make this comparison for the confidence of the reader. The RMA normalisation was completed using Partek®, and is therefore discussed in detail in section 2.1.3.4.

### 2.1.3 Statistical analysis

To compare changes in gene expression profiles in different myofibroblasts populations, three datasets were compiled for analysis, CAM vs. AMN, CAM vs. ATM and ATM vs. AMN.

#### 2.1.3.1 Background oligonucleotides

As a specific cell type may not produce every single known protein, we first compiled lists of genes that were expressed in each cell-type. Within this study, all genes found to be expressed at a baseline level were pooled to define the list of expressed genes for each of the myofibroblast populations. For these studies an oligonucleotide was only classified as being detected, if a P flag was present in 100% of patients from either or both of the comparison groups.

#### 2.1.3.2 Differentially regulated oligonucleotide lists.

Microarray data was processed through GeneSpring GX.10 software (Agilent). As myofibroblasts were prepared from paired cancer and adjacent samples, a paired T-test was conducted on data from these samples. As patient's sz45, sz268, sz294 and sz305 had a single CAM sample but two ATM samples, two independent paired T-tests were performed using the same CAM sample with each of the two ATM samples. Therefore, in total 15-paired t-tests were performed. No genes passed a p-value ≤0.05 Benjamini-Hochberg false discovery rate FDR correction, this may be due to the samples being paired and therefore very similar. Therefore, in each case

oligonucleotides with un-corrected p-values ≤0.05 were considered to be differentially expressed. Un-paired T-tests were performed on array data derived from CAM vs. ANM and ATM vs. ANM comparisons. In each case P-values were corrected for false discoveries in accordance to the Benjamini-Hochberg false discovery rate (FDR) algorithm. FDR corrected p-values ≤0.05 were considered differentially expressed. To determine if genes were over or under-expressed, the log fold change was calculated for each oligonucleotide and the geometric mean was calculated for oligonucleotide intensities within each sample group, i.e. CAM, ATM or ANM. For each differentially expressed oligonucleotide in the CAM vs. ATM, CAM vs. ANM and ATM vs. ANM studies a log base 2 transformation was applied to ensure that the relative changes in the data are proportionate, with changes around negative, or small intensities being comparable to changes around large intensities.

#### 2.1.3.3 Patient prognosis groups

'Good' and 'bad' patient prognosis groups were compiled from the four best and four worst prognosis scores respectively, with the intention of identifying changes in myofibroblasts derived from patients with a worse or better prognosis score. In addition to sub-grouping patients into 'good' and 'bad' prognosis groups, we also wanted to isolate changes that occur in myofibroblasts derived from the cancer or adjacent regions, in order to highlight the stage (CAM or ATM) that changes occurred during tumour development.

##### 2.1.3.3.1 'Good' patient prognosis

'Good' patient prognosis scores are classed as those below 9, therefore the four patients with the best prognosis scores were sz42, sz194, sz198 and sz389, which

relate to samples 1, 5, 7 and 15, respectively. Un-paired T-tests were conducted on the 4 'good' CAM samples vs. all 12 ANM samples, or the 4 'good' ATM samples vs. all 12 ANM samples (Table 2.6). In each case, P-values were corrected for false discoveries using the Benjamini-Hochberg correction algorithm. Corrected p-values ≤0.05 were considered to be differentially expressed and the $\log_2$ fold-change was calculated for each oligonucleotide.

| 'good' patient prognosis groups | |
|---|---|
| CAM vs. ANM | |
| 1-CAM | All 12 Absolute normal samples |
| 5-CAM | |
| 7-CAM | |
| 15-CAM | |
| ATM vs. ANM | |
| 1-ATM | All 12 Absolute normal samples |
| 5-AMN | |
| 7-ATM | |
| 15-AMN | |

**Table 2.6** 'good' patient subgroups, CAM vs. ANM and ATM vs. ANM.

#### 2.1.3.3.2 'Bad' patient prognosis

Patients with prognosis scores above 9 were classified as having 'bad' prognosis, therefore the four patients with the worst prognosis score were sz190, sz192, sz271 and sz308, which relate to samples 3, 4, 9 and 12, respectively. Un-paired T-tests were conducted on the 4 'bad' cancer samples vs. all 12 absolute normal samples and the 4 'bad' adjacent samples vs. all 12 absolute normal samples (Table 2.7). P-values were again corrected for false discoveries using the Benjamini-Hochberg correction algorithm described in section 1.1.4.2. All corrected p-values ≤0.05 were considered to be differentially expressed and $\log_2$ fold change values were calculated for each oligonucleotide.

| 'Bad' patient prognosis group | |
|---|---|
| CAM vs. ANM | |
| 3-CAM | All 12 absolute normal samples |
| 4-CAM | |
| 9-CAM | |
| 12-CAM | |
| ATM vs. ANM | |
| 3-ATM | All 12 absolute normal samples |
| 4-ATM | |
| 8-AMN | |
| 12-ATM | |

**Table 2.7** 'Good' patient subgroups, CAM vs. ANM and ATM vs. ANM.

#### 2.1.3.4 Partek®

The Partek® Genomics suite, version 6.6 beta Copyright © 2008 (Partek® Inc., St. Louis, MO, USA) became available to us towards the latter part of this project. Throughout the study, data was normalised using Mas5, as described in section 2.1.2, however due to the increasing popularity of an alternative normalisation technique, Robust Multi-array (RMA), it seemed sensible to compare the two. Affymetrix CEL files were uploaded into Partek® and normalised using the RMA algorithm. In brief; background correction on the PM values, quantile normalisation across all chips on experiment, Log2 transformation and a median polish summarisation. The resulting RMA normalised expressions were exported from Partek and differentially resulted gene lists and their associated fold changes calculated as described in section 2.1.3.2.

Partek® was also used to assess the quality of the data and ensure correct fold changes were applied. Briefly, the data was transposed and patient IDs were identified as categorical factors and oligonucleotide expression levels as double

response factors. Then additional information was added manually; Individual patient samples were annotated based on their sample type (cancer, adjacent or normal), any patient pair information for cancer and adjacent samples and batch processing information were taken into account.

Principal component analysis (PCA) was performed on un-corrected data, data corrected based on microarray analysis dates, and data corrected on patient pair information. In each case, results were visualised as 2 and 3 dimensional plots. Data corrected based on microarray analysis dates takes into account any systematic variation that may be associated with sample preparation, whilst correction based on patient pair samples, takes into account variation between individuals. Also within Partek®, differentially expressed gene lists for each comparative dataset were identified using the ANOVA method. Factors within the ANOVA included sample type and for the CAM vs. ATM dataset, patient pair information. Ranges of differentially regulated gene lists were compiled based on different stringencies as stated below. Differentially regulated genes had to have a p-value $\leq 0.05$ and have fold changes greater than 2, 1.8, 1.6, 1.4, 1.2 or 1 fold change. As before, p-values were corrected using the Benjamini-Hochberg false discovery rate for the CAM vs.. ANM and the ATM vs. ANM dataset, but not for the CAM vs. ATM dataset. Differentially expressed gene lists were visualised using hierarchical clustering analysis (HCA) within Partek® at each of the fold change thresholds, thereby allowing the best fold-change cut-off to be selected.

## 2.2 Metacore™

Metacore™ (GeneGo Inc) is a commercial integrated software suite for network and pathway analysis. Metacore™ allows different data-types to be uploaded, visualised and analysed with the use of its manually curetted knowledge database of protein-protein interactions, transcription factors, drug targets, metabolic pathways and signalling pathways.

### 2.2.1 Conversion of oligonucleotide probe IDs to Genes

For each dataset, text files of Affymetrix oligonucleotide associated log fold-changes and p-values were imported into Metacore™. For each oligonucleotide, associated Entrez gene Ids were exported from Metacore™. For genes that are represented by multiple oligonucleotide probes, Metacore™ selects the highest probe expression value.

### 2.2.2 GeneGo Pathway analysis

For each data set, significantly changed Entrez gene IDs and their associated background detection list were uploaded into Metacore™. Individually, a significantly changed gene list was activated and the appropriate background list set as the 'universe'. The Enrichment ontology's - GeneGo pathway function was then applied to identify over-represented pathways by calculating if there are statistically, more significantly changed genes present in a particular pathway than would be expected by random chance. Pathways are also given a False Discovery Rate (FDR) corrected p-value. The FDR correction takes into account the large number of Fisher tests applied to the datasets and calculates the chance of incorrectly rejecting the null hypothesis.

### 2.2.3 Retrieval of assigned gene lists

Lists of significantly changed genes mapped to a Metacore™ GeneGo pathways was obtained through 'File > properties > on maps'. This displays the number and which significantly changed genes are mapped to a pathway within Metacore™. Retrieving lists of significantly changed genes mapped to significantly over-represented pathways was more problematic, as pathway member information for a select group of pathways cannot be exported from within enrichment ontology's - GeneGo pathway results.

Using 'View > GeneGo maps' (completed during 2010, this function has currently expired), all pathways that had at least 1 significantly changed gene hit are listed and groups of pathways exportable, however no associated pathway over-representation p-values are displayed. Therefore, this list had to be cross-referenced with the over-represented p-values obtained from enrichment ontology's - GeneGo pathway results. Pathways within 'View>GeneGo maps', were only selected if their over-represented p-value ≤0.05.

### 2.2.4 Transcription factor analysis

Metacore™ was used to identify transcription factors within differentially expressed gene lists. Two lists of differentially regulated transcription factors were compiled for each dataset, those with a p-value ≤0.05 and those with a p-value ≤0.05 and a 1.6 fold change cut-off. These two lists of transcription factors were then identified in the differentially expressed gene lists in Partek® and transcription factor heatmaps generated for the CAM vs. ANM, CAM vs. ATM and ATM vs. ANM datasets.

## 2.3 DAVID

The Database for Annotation, Visualization and Integrated Discovery (DAVID) was used to assess over-represented KEGG and BioCarta pathways within the datasets (Athippozhy et al., 2011; Huang et al., 2009; Huang et al., 2008). Individually, lists of significantly changed Entrez gene IDs and their associated background Entrez IDs were uploaded into the DAVID bioinformatics resource. Numbers of genes mapped within the DAVID database were recorded. The functional annotation chart tool within DAVID allows significantly changed genes to be mapped onto related biological pathways, specifically KEGG and BioCarta. Datasets were processed through the functional annotation tool: In this study, all pre-defined defaults were cleared, KEGG and BioCarta Pathways were selected and a functional annotation chart analysis was performed on the data. Functional annotation chart results provide all related biological KEGG and BioCarta pathways, with associated over-representation pathway p-values and Bonferroni- Hochberg FDR corrected p-values. Pathways with FDR corrected p-values ≤0.05 were considered over-represented in the significantly changed gene-list.

### 2.3.1 Retrieval of assigned gene lists

To retrieve the total number of significantly changed genes that were mapped to KEGG and BioCarta pathways in DAVID, genes that were not mapped to pathways were downloaded via the functional annotation chart before being deducted from our total significantly changed gene list. Retrieval of significantly changed genes mapped to significantly over-represented pathways ($p \leq 0.05$) was obtained by

selecting and downloading the 'genes' column in the functional annotation chart results for each significantly over-represented pathway.

## 2.4 Ingenuity®

IPA 7 is part of Ingenuity® systems (Ingenuity® Systems, www.Ingenuity.com), a commercial web-based software database. IPA 7 allows the analysis of microarray, micro-RNA and proteomics data, which can be used to study gene, protein and drug information. Ingenuity's knowledge database is manually curated by PhD level scientists reading full-text articles. Datasets are mapped onto the Ingenuity® knowledge database of canonical metabolic and signalling pathways, derived from papers, reviews, textbooks and the KEGG database. Ingenuity® allows users to either upload their own reference set or apply the Ingenuity® knowledge database as a reference set. A combined list of background and significantly changed genes along with their associated p-values were uploaded into Ingenuity®. If p-values had not been calculated for background genes, a standard p-value of 0.99 was assigned. Once uploaded the following filters and settings were assigned: The background was set as 'user defined' (this is the entire list of background and significantly changed genes uploaded), direct and in-direct relationships, all species, all tissues and all data sources were selected. To select for significantly changed genes a cut-off p-value of 0.05 was applied. Lists of significantly changed genes, that were mapped to different Ingenuity® categories, were exported as: mapped to Ingenuity®, mapped to canonical pathways, mapped to networks, and not mapped. The pathway analysis was set to run; Exported canonical pathway results display: pathways, -log (p-value), p-value (un-corrected or corrected), ratio and significantly

changed genes hitting a specific pathway. The ratio is a calculation of the number of significantly changed genes in a pathway divided by the total number of genes in a pathway. This provides information about which pathways have the most significantly changed genes in them and therefore may be most affected by differentially expressed genes. P-values are calculated using the Fisher exact test and P-values were displayed as $-\log_{10}$ values, therefore –log values ≥ 1.3 are equivalent to p-values ≤ 0.05. Pathways with p-values ≤ 0.05 were considered to be over-represented. FDR corrected p-values for over-represented pathways were obtained. Settings were also altered to display FDR corrected over-represented p-values based on Benjamini Hochberg FDR algorithm. JPEG images and lists of over-represented canonical pathway results were exported for un-corrected and FDR corrected p-values separately.

### 2.4.1 Retrieval of assigned gene lists

The list of genes mapped to Ingenuity® pathways shows the input Entrez IDs and their associated gene symbols within the Ingenuity® knowledge database. All other lists of genes exported from Ingenuity® are in the form of official gene symbols rather than Entrez gene IDs. Therefore, the 'mapped to Ingenuity®' gene-list conversions together with the VLOOKUP function in excel was used to convert all gene symbols exported from Ingenuity®, to the required Entrez gene IDs. Lists of significantly changed genes mapped to statistically over-represented pathways were compiled by merging significantly changed genes mapped to all pathways with p-values ≤ 0.05, within the canonical pathway results table. Lists of significantly changed genes mapped to any canonical pathway within Ingenuity® was obtained

by merging all significantly changed genes mapped to all pathways from the canonical pathway results table. The canonical pathway results table only includes pathways with p-values large enough to visualise on chart (as depicted by blue bars on the canonical pathway charts). Therefore, to obtain a complete list of genes hitting all pathways regardless of resultant pathway over-representation p-value, additional pathways need to be individually selected and significantly changed genes that are mapped to it exported.

## 2.5 Reactome

Reactome is an online open-resource interaction, reaction and pathway database (Matthews et al., 2009; Vastrik et al., 2007). Its database is manually curated and peer-reviewed, from NCBI, PubMed, Uniprot, Entrez gene ID, ensemble, KEGG and Gene Ontology. The Reactome database is freely available to download and Reactome data model consists of individual reactions being grouped into networks, of which entities interact, to form larger biological pathways. Reactome Skypainter (Feb 2010) allows users to upload experimental data, and through the application of a one-tailed Fisher exact test, determine which pathways are statistically over-represented in the dataset. However, Reactome Skypainter does not allow the application of a custom background, therefore applies the entire Reactome database as the background for statistical pathway over-representation calculations. Neither does Reactome apply a false discovery rate correction for the large amount of Fisher exact tests being carried out.

### 2.5.1 R computing language

The R computing language is a statistical and graphical computing software environment based on the S programming language, created by John Chambers and Bell Labs. Bioconductor provides a range of tools, which can be integrated with R for the statistical analysis of genome-wide studies (Gentleman et al., 2004). The Bioconductor package BioMart (Durinck et al., 2005) was used to access the Reactome pathway database for statistical analysis in R. Full details of R script can be found in Supplementary script 1.

#### 2.5.1.1 Array Quality Metrics

The Bioconductor package ArrayQualityMetrics was used to assess the relative quality of the different gene arrays within the dataset. Identification of variance and noise due to technical causes such as the platform, lab, experimentalist, RNA extraction, amplification, hybridisation, labelling and sampling need to be assessed to ensure biological effects are visualised and statistical power is not reduced. ArrayQualityMetrics was applied before normalisation, to all 39 raw Affymetrix CEL files. The ArrayQualityMetrics report can be subdivided into the following sections: individual array quality; comparison of homogeneity between arrays; variance mean dependency and Affymetrix specific plots. Full details of R script can be found in Supplementary script 2.

#### 2.5.1.2 Pathway analysis and BioMart assigned gene lists

The format of the output results from the Reactome over-representation pathway script, are Reactome unique pathway IDs. I wrote an R script to extract Reactome unique pathway IDs and their related Pathway names from the Bioconducter

software BioMart. Reactome unique pathways were converted to pathway names to aid biological interpretation of the pathway results. Full details provided in Supplementary script 3.

## 2.6 Cytoscape

Cytoscape is an open source software platform for the integration and visualisation of molecular interaction networks (Cline et al., 2007). It was created at the institute of systems biology, in Seattle in 2002. Either version 2.7 or 2.8 have been used to generate protein-protein interaction network figures, proteins are represented by nodes and interactions represented by edges.

### 2.6.1 Human protein interaction network

In December 2008, Dr. R. Hyde (Sanderson Lab, University of Liverpool) constructed a human and interolog protein-protein interaction network, as previously published (Markson et al., 2009). This was based on the collation of many different human and interolog protein interaction databases, as yet a database containing all known human protein interactions does not exist. The databases are comprised of manual literature curation by expert biologists and direct user submissions, all are freely downloadable. An interolog is a predicated conserved interaction between two proteins that have orthologs that interact in another species. These predicted interactions may provide insight into the function of currently un-characterised human proteins.

Four human databases were used to collate the human protein interaction data; Molecular INTeraction Database (MINT) (Ceol et al., 2010), The Human Protein Reference Database (HPRD) (Keshava Prasad et al., 2009), BioGrid (Breitkreutz et al.,

2008) and IntAct (Kerrien et al., 2007). Interolog interaction data was collated for *Drosophila melanogaster* (*D. mel*), *Caenorhabditis elegans* (*C. ele*), *Mus musculus* (*M. mus*) and *Saccharomyces cerevisiae* (*S. cerevisae*) from MINT, BioGrid, IntAct with orthology calls from the NCBI tool-homologene (Sayers et al., 2009) and inParanoid (O'Brien et al., 2005). Combining all mined data resulted in a human and interolog protein-protein interaction network with 13,101 nodes (proteins) and 143,179 edges (interactions). All gene identifiers were converted to Entrez Gene Ids for standardisation throughout the analysis.

### 2.6.2 Myofibroblast expression networks

As all genes are not expressed in all cell types, it was important to generate a myofibroblast specific gene expression network, which contained all genes detected in this specific cell-type at a baseline level. To generate this network, an attribute file was produced containing a union of genes detected at baseline level (background) in CAM vs. ANM, CAM vs. ATM and ATM vs. ANM datasets. Attribute files were assembled for each experimental group, CAM vs. ATM, ATM vs. ANM and CAM vs. AN. Attribute files contained either differentially expressed Entrez gene IDs and Log fold changes or differentially expressed Entrez gene IDs and p-values. P-values are un-corrected for the CAM vs. ATM dataset, and FDR corrected for the CAM vs. ANM and ATM vs. ANM. The myofibroblast network was opened in Cytoscape, attribute files for each comparison group were imported and selected using 'old filters'. New networks of differentially expressed genes were then generated for each comparison group, with nodes being assigned colours according to the direction of log-fold change.

#### 2.6.2.1 'Good' and 'bad' myofibroblast expression network

As the 'good' and 'bad' patient sub-groups represent data from 4 cancer and 4 adjacent samples, rather than the full 12, a specific gene expression network was also generated for this sub-set of myofibroblasts. The resulting 'good' and 'bad' gene expression network was produced by first compiling an attribute file containing a union of genes detected at a baseline level (background) of all the 'good' and 'bad' patient prognosis backgrounds, for the CAM vs. ANM and the ATM vs. ANM datasets. The 'good' and 'bad' gene expression network was then produced exactly as the standard myofibroblast gene expression network, described in section 2.62. Attribute files were assembled for each experimental group, CAM vs. ANM, 'good' and 'bad' and ATM vs. ANM 'good' and 'bad'. Attribute files contained either differentially expressed Entrez gene IDs and Log fold-changes, or differentially expressed Entrez gene IDs and p-values. Subsequent analysis using Netbox required 'good' and 'bad' attribute files to be set up using official gene names, therefore extra attribute files were generated, containing either differentially expressed official gene names and log fold-changes, or differentially expressed official gene names and p-values. All p-values used for 'good' and 'bad' attribute files were FDR corrected. The myofibroblast network was opened in Cytoscape and attribute files for each comparison group were imported, selected uses 'old filters'. New networks of differentially expressed genes were generated for each comparison group, with nodes being colour coded according to the direction of Log fold change.

### 2.6.3 Hypernode

Hypernode is a Cytoscape plugin designed by Dr. Russell Hyde (University of Liverpool). Utilising the KEGG pathway database, all human pathways and gene members were downloaded and utilised within the hypernode plugin. It is able to condense pathway genes on to 'hypernodes' if they are present in a single pathway, genes which are present in >1 pathway appear as nodes connecting multiple 'hypernodes'. Datasets were uploaded into Cytoscape and visualised using the plugin, allowing the complex connectivity between pathways and highlighting genes providing the crosstalk to be visualised.

### 2.6.4 BiNGO

The Biological Network Gene ontology tool BiNGO™ (Maere et al., 2005) is a Cytoscape plugin that allows networks to be analysed for the over-representation of Gene Ontology (GO) annotations. Lists of genes or networks can be analysed, and visualised within Cytoscape as a map of hierarchical GO annotations. BiNGO™ results also contain a GO annotation text file, with details of input genes linked to over-represented GO annotations. To calculate which GO terms were over-represented in the differentially expressed gene set, BiNGO™ requires a background gene list to be set. For this purpose background gene lists from individual datasets were assembled and uploaded into Cytoscape, enabling differentially expressed genes to be identified from relevant background networks prior to performing BiNGO™ analyses. The hypergeometric test was used to look for over-representation of GO annotations, in differentially expressed gene sets compared to their individual backgrounds. For each dataset the analysis was run twice, firstly to

visualise all GO annotations with un-corrected p-values ≤0.05 and then for the visualisation of GO annotations with Benjamini-Hochberg FDR corrected p-values ≤0.05. Individually, gene lists were compiled of Entrez gene IDs falling within Go annotations with p-values ≤0.05 and GO annotations with FDR corrected p-values ≤0.05 for further analysis.

## 2.7 Comparison of canonical pathway analysis tools.

As initial pathway analysis was conducted in Metacore™, for each dataset, lists of differentially regulated genes un-mapped to pathways in Metacore™ were assembled. Analysis of these Metacore™ unmapped genes was then performed using DAVID, Reactome and Ingenuity® to provide an indication of the distinct advantages that each analysis tool may provide. There was a selection of genes that could not be mapped to a canonical pathway by any of the four tools. BiNGO's GO annotations, was finally used to try to provide further insight into the biological roles of these genes.

## 2.8 Epigenetics

A list of chromatin remodelling factors was manually curated from the Human Epigenetic Chromatin Remodelling factors $RT^2$ Profiler™ PCR Array (SABiosciences, QIAGEN, Frederick, USA). These genes were combined with a list of epigenetic modification enzymes extracted from the Human Epigenetic Chromatin Modification Enzymes $RT^2$ Profiler™ PCR Array (SABiosciences, QIAGEN, Frederick, USA) and a list of proteins involved in epi-microRNAs, epigenetic genes which are regulated by microRNAS (Iorio et al., 2010). Gene names were converted into Entrez gene IDs using the DAVID bioinformatic gene ID conversion tool. From this,

epigenetic attribute files were uploaded into Cytoscape and selected from the entire Human interactome with interologs, resulting in an epigenetic network, allowing for the mapping of experimental data and clear visualisation of epigenetic changes.

#### 2.8.1.1 'Good' and 'bad' prognosis epigenetic networks

Entrez gene ID attribute files compiled in section 2.6.2.1 for 'good' and 'bad' patient's differentially regulated genes were mapped onto both epigenetic networks. Colours of nodes represent the direction of the fold change. Within each experimental group, CAM vs. ANM and ATM vs. ANM, epigenetic changes seen in the 'good' and 'bad' patients were exported from Cytoscape and unique and common epigenetic changes for the 'good' and 'bad' patient prognosis groups was recorded.

## 2.9 Multivariate analysis

Multivariate analysis (MVA) was carried out with the help of statistician (Richard Jackson, University of Liverpool) using the R computing language environment. Full details of Fisher tests, matrix formation and MVA similarities based on genes or based on pathways, and correspondence analysis are provided in Supplementary scripts 4-8 respectively.

### 2.9.1 Pathway over-representation tests

Previously mentioned pathway analysis tools use the Fisher exact test to determine whether a pathway is statistically over-represented within a significantly changed gene set. In this study three different techniques were used to detect over-represented pathways, or processes, within a given gene list. These include the

Fisher exact test, and Chi-Squared tests. Below is an example of the 3x2 contingency table formed to carry out such statistical tests. In contrast, a 2x2 contingency table was formed using only a, b, d and e, not taking into account genes not in the background set.

| | Pathway | |
|---|---|---|
| | Yes | No |
| Diff. Expressed | a | d |
| Background | b | e |
| Not in background | c | f |

The odds ratio was thought to be a more appropriate scoring system to apply to define over-represented pathways;

$$\text{odds ratio} = \frac{a*e}{b*d}$$

The odds ratio takes into account the proportion of differentially expressed genes in a pathway to the number of non-differentially expressed (background) genes in a pathway. Therefore, if an odds ratio is greater than 1 the proportion of differentially expressed genes in a pathway is greater than the proportion of background genes in a pathway.

### 2.9.2 Matrix formation

Multivariate analysis uses a dissimilarity matrix of 1's and 0's; in our case, the numbers identify which genes are present in which pathways. Within R, Reactome

pathway mappings from pathway ID to Entrez gene ID was extracted from BioMart, though the Bioconductor website. Significantly changed gene lists and background gene lists were read into R, and only those genes and associated pathways extracted from Reactome with genes in our background lists were used in the analysis. The matrix was then calculated, pathways form the matrix columns and genes the rows, the script cycles through each gene, placing a '1' if the gene is present in the given pathway and '0' if it is absent.

### 2.9.3 Multivariate analysis – similarities of genes

Similarity of genes, calculates the similarities of genes based on the pathways in which they are present. The genes/pathways matrix includes all background and significantly changed genes. To reduce the amount of noise the matrix was trimmed to remove all genes that were present in less than two pathways and all pathways that had less than two significantly changed genes were also removed. Using a binary method, the trimmed matrix was then converted into a matrix of distances, based on how different the genes are to one another. A value of zero represents no difference between two genes and a value of 1 represents two genes which are identical to each other (with regards to the pathways they appear in). Hierarchal cluster analysis (HCA) was then conducted using the average link algorithm on the dissimilarity matrix. The method initially tries grouping the most similar genes into pairs, after which its groups the next pair of similar genes and so on until all genes are connected in a tree-like dendrogram. Tight clusters of genes represent a similar pathway representation.

Multidimensional scaling (MDS) also uses the dissimilarity matrix, the MDS object is created which provides a set of MDS co-ordinates to be plotted in a 2 or 3-dimensional space. Points on the plot represent genes, while geometrical distances provide a measure of dissimilarity between one another. Red points represent significantly changed genes and black nodes represent background genes. Significantly changed genes in clusters or with similar geographical positions, were grouped and exported from R for further analysis. Reactome pathway information for each group of genes was exported from the BioMart website.

### 2.9.4 Multivariate analysis - similarities of pathways

Similarity of pathways, calculates the similarities of pathways based on the genes that they include. The genes/pathways matrix only includes significantly changed genes. To reduce the amount of noise and identify real groups of pathways within the data the matrix was trimmed to remove all pathways that contain less than two significantly changed genes and all genes present in less than two pathways. Using a binary method, the trimmed matrix was converted into a matrix of distances, based on how different pathways are to one another. A value of zero represents no difference between two pathways and a value of 1 represents two pathways, which are identical to each other (with regards to the genes they contain). Hierarchal cluster analysis (HCA) was conducted using the average link algorithm on the dissimilarity matrix. HCA groups pathways using the same method described in 'similarity based on genes'. Tight clusters represent pathways with a similar gene makeup.

Multidimensional scaling (MDS) also uses the dissimilarity matrix, the MDS object is created which provides a set of MDS co-ordinates to be plotted in a 3-dimensional space. Points on the plot represent pathways and their geometrical distances will represent a measure of dissimilarity between one and other. Red points on the plot represent pathways with an odds ratio of 2 or greater, which are classed as over-represented and black nodes represent pathways with an odds ratio less than 2. Over-represented pathways that form clusters/have a similar geographical position were identified and exported from R for further analysis. Entrez gene ID information relating to each group of pathways was then obtained from the BioMart website.

### 2.9.5 Correspondence analysis

Correspondence analysis is a multi-variant analysis technique, which is similar to multi-dimensional scaling, in that they both plot the difference between observation in a point in space. The difference is, multi-dimensional scaling plots genes or pathways, whereas, correspondence analysis allows Genes and Pathways to be considered at the same time and plots them both on the same set of axis. In this approach, clusters not only identify interesting genes but also associated pathways.

As previously described in MVA analysis section 2.10.1, correspondence analysis uses an odds ratio to define which pathways are over-represented and the compilation of a pathway and gene matrix. The genes/pathways matrix includes all background and significantly changed genes. To reduce the amount of noise the matrix was trimmed to remove all genes that were present in less than two pathways and remove all pathways that had less than two significantly changed

genes in them. Correspondence analysis provides a set of co-ordinates to be plotted in a 3-dimensional space, points on the plot represent genes (red) and pathways (black) and their geometrical distances provides a measure of dissimilarity between one another. Central dense clusters were exported from R for further analysis, as points of potential high connectivity or cross-talk.

## 2.10 Netbox

Netbox is a java-based online resource, which allows the input of data onto a human interaction network, and use of methods to identify unique linker genes, network modules and the generation of random modules to statistically analyse global and local module connectivity (Cerami et al., 2010). They produced a human interaction network, curated from literature resources only. The human interaction network includes protein-protein interactions and signalling pathways from four databases, HPRD (Keshava Prasad et al., 2009), Reactome (Matthews et al., 2009), HCI-Nature pathway interactome database (PID) (Schaefer et al., 2009) and MSKCC cancer cell map.

Linker genes are genes that themselves are not differentially expressed but which link two differentially expressed genes together. Statistically significant linker genes were identified by their global degree and their hypergeometric distribution, also known as the one-tailed Fisher exact test. The Benjamini Hochberg FDR correction is applied and those passing a P-value $\leq 0.05$ are considered statistically significant. The edge betweenness algorithm (Girvan and Newman, 2002) was used to detect modules within the network. The betweenness algorithm was first introduced in 1977 (Freeman, 1977), as a measure of information flow through a network; an

edge lying on a large number of shortest paths has a large control over the information flow of a network. This process is used to detect communities within networks, an edge with a high betweenness value may lie on a large number of shortest paths but have a low degree and therefore it may connect two areas of highly connected nodes or modules. The edges betweenness algorithm used here removes edges with a large betweenness value from the network successively. Modules are assigned a modularity score, using an assortive mixing algorithm (Newman, 2003). If the number of edges within the module is no better than expected by random chance then the modularity score will approach 0, while a modularity score of 1 indicates a highly modular network.

To assess the statistical significance of the networks modularity score, random simulations are run on a network of the same size, same number of nodes and degree distribution but a random selection of interaction partners. The time (O) the algorithm takes to run is based on the formula: O (mn) (Newman, 2001), were m is the number of edges and n is the number of nodes in the network. The algorithm needs to be re-run after every edge is removed, so a worst time of O ($m^2n$) is predicted. As this algorithm is immensely time consuming, the need to produce 1000 random simulations becomes impractical, instead 100 random simulations are run. The modularity score of the network is then altered to take into account the random re-wired network and a scaled modularity score given, known as the z score. The z score is calculated by taking the modularity score for the actual network, deducting the modularity score for the random network and then dividing

by the standard deviation of the random re-wired network. (Wang and Zhang, 2007).

### 2.10.1 Settings and command script

Netbox was downloaded from http://cbio.mskcc.org/tools/netbox.html, unzipped and added as an environment variable to the Netbox directory. Configuration files were set up for CAM vs. ANM, CAM vs. ATM, ATM vs. ANM, CAM vs. ANM ('good' and 'bad' separately), and ATM vs. ANM ('good' and 'bad' separately). Within configuration files, the shortest path threshold was set as 2, thus allowing differentially regulated genes to be indirectly connected by statistically significant linker genes. The p-value threshold for linker genes to be added to the network was ≤ 0.05.

The following code was run in the computers command prompt:

*C:\> cd netbox*

*C:\netbox> cd gbm_data*

*C:\netbox\gbm_data C:\Python27\python.exe c:\netbox\bin\netAnalyze.py 'name of configuration file' –i*

The *–i* allows the linker p-value to be varied within the configuration file, before producing a network. Network sif files were uploaded into Cytoscape, modules were genes categorised into a certain module are identified using modulues.txt attribute files and linker nodes distinguishable from differentiated genes by uploading the node_type.txt attribute files.

## 2.11 Identification of previous un-assigned genes

After pathway analysis, in Metacore™, DAVID, Reactome, and Ingenuity® many differentially regulated genes are still unassigned to canonical pathways. As there is no pathway information for these unmapped genes, we wanted to try to understand what these genes are involved in, and how they relate to the significantly over-represented pathways. Although unmapped genes are not mapped to canonical pathways we screened the NetBox network to see if any unmapped genes had been assigned to modules or connect modules by interacting with members of multiple modules. This is important to our understanding of the roles the differentiated unmapped genes may play, as modules are believed to represent functional modules/biological organisation within the cell (Hartwell et al., 1999).

## 2.12 Correlation analysis

Through collaboration with statistician Richard Jackson (University of Liverpool), we were able to perform correlation analyses to look at the relationship between the fold change of genes and patient prognosis scores. Outside of R, for the un-paired datasets, CAM vs. ANM and ATM vs. ANM, the arithmetic mean of all absolute normal samples was calculated and this value was used to calculate individual patient log fold changes. Within R, correlation analysis was performed using Supplementary scripts 9-11, a separate script for each of the datasets. CAM vs. AN, ATM vs. ANM and CAM vs. ATM datasets of individual patient logged intensities were opened in R. Individual patient log fold-changes were calculated for each Entrez gene ID. As multiple probes map to a single gene, there are a select number

of duplicate genes in the dataset, the average log FC was then taken for such genes. Patient prognosis scores are read into R and two different types of correlation analysis was calculated. Spearman's correlation uses the ranked prognosis scores and Pearson's correlation uses the actual prognosis scores. Genes that are highly correlated with prognosis have values close to +/-1, which represent genes that are either positively or negatively correlated with prognosis. In this analysis, a positive correlation represents a gene with a larger fold change in patients with worse prognosis scores, whilst a negative correlation represents a gene with a smaller fold change in patients with worse prognosis scores. Data was visualised as heapmaps generated in R. A series of different correlation analysis thresholds were applied to select the optimum correlation threshold for each dataset. In addition, a box plot was generated for each dataset within R, in order to display the variance of the log fold-changes for each gene for individual patients and a plot of prognosis score vs. log fold change for every gene and patient.

## 2.13 T-Test

Through collaboration with statistician Richard Jackson (University of Liverpool), genes that have statistically significant difference in expression levels between 'good' and 'bad' patient sub-groups were identified. Outside of R, as the 'good' and 'bad' sets are generated from the un-paired CAM vs. ANM dataset, as before, the arithmetic mean of all absolute normal samples was calculated and this value was used to calculate each individual patient log fold-change.

Within R, a t-test was performed to test the statistical difference between the 'good' and 'bad' log fold-changes for each gene was applied using the script

provided in Supplementary script 12. The logged intensity values of all cancer samples for the 'good' and 'bad' sets and absolute normal patient arithmetic means were opened in R. Individual patient log fold-changes were calculated for each Entrez gene ID. As multiple probes map to a single gene, there are a select number of duplicate genes in the dataset, the average log fold-change was then taken for such genes. The individual patient samples were annotated as 'good' or 'bad' and a T-Test was run to identify genes with statistically different fold change in expression in the 'bad' and 'good' sets. Lists of genes with statistically different expression profiles were assembled at three different p-value thresholds, $p \leq 0.05$, $p \leq 0.01$ and $p \leq 0.005$ and visualised using heat maps.

# 3 Chapter Three: Comparative gene expression profiling

## 3.1 Introduction

The primary aim of this study was to develop a better understanding of how myofibroblasts derived from the site of gastric tumours or adjacent tissue may be changed relative to normal gastric myofibroblasts. By defining changes in gene expression patterns for myofibroblasts derived from a range of patients with varying prognostic scores we aimed to identify gene signatures, which may relate to different stages of tumour development and/or 'good' vs. 'bad' patient prognostic groups. In addition, we wished to use pathway enrichment methods to provide insight into the molecular processes that are changed in CAMs as opposed to ANMs.

## 3.2 Isolation and characterisation of primary myofibroblast cell lines

Details of the isolation and characterization of primary human gastric myofibroblasts that were used to generate microarray data for this study are described in Holmberg *et al.* (2012). In brief, human primary myofibroblasts were isolated from resected gastric tumours (CAMs) or matched adjacent tissue (ATMs) by Dr Peter Hegyi, (Department of Medicine, University of Szeged, Hungary), following protocols described by (Wu et al., 1999). Absolute Normal Myofibroblasts (ANMs) were isolated from deceased transplant donors with normal gastric morphology. Figure 3.1 (taken from Holmberg et al 2012) shows examples of changes in myofibroblast numbers and tissue architecture in tissues from which CAMs, ATMs or ANMs were derived. To ensure that all tissues from which CAM, ATM and ANM samples were purified had expected and consistent tissue profiles

myofibroblast morphology, abundance and tissue architecture were quantified (Figure 3.1D).

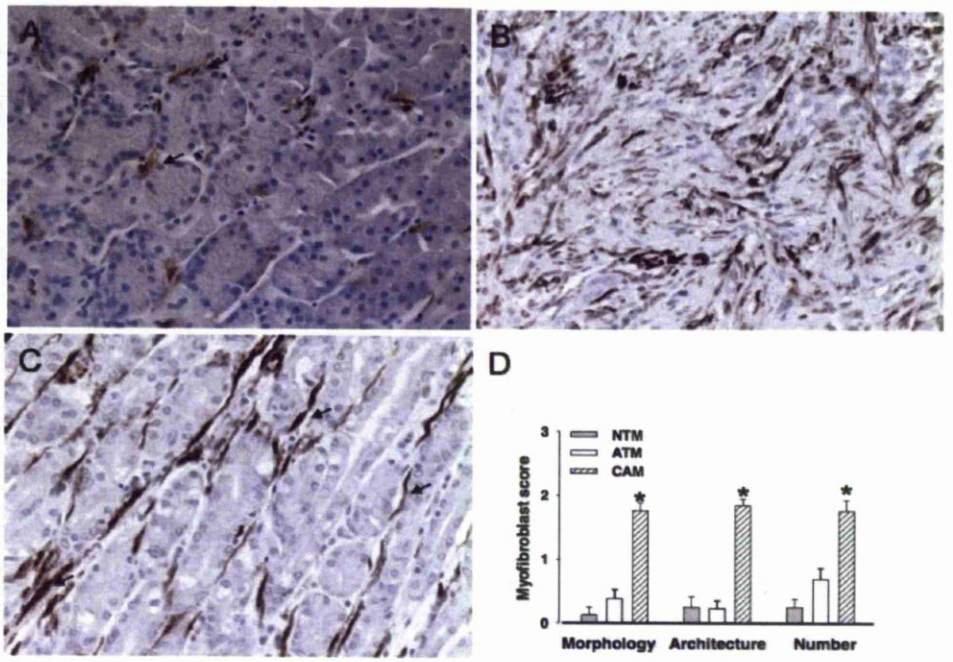



**Figure 3.1** Representative images of tissues from which ANMs (A), CAMs (B) or ATMs (C) were isolated. Brown stain shows localization and relative abundance of myofibroblasts in each tissue. D) Quantification of myofibroblast morphology, architecture and number in cancer, adjacent and normal tissues. * ANOVA, $p<0.05$ CAM vs. ATM, CAM vs. ANM. From Holmberg et. al. Carcinogenesis 33:1553 (2012)

All tissue samples were scored for myofibroblast morphology (0 = normal, 1 = mildly distorted or 2 = severely distorted) and tissue architecture (0 = myofibroblasts restricted to periglandular or subepithelial localisation, 1 = myofibroblasts located in periglandular and sub-epithelial regions and elsewhere in the interstitium or 2 = severe architectural damage with meshwork-like appearance). In addition, myofibroblast numbers were also scored as 0: similar to normal tissue, 1: mild to moderately increased or 2: substantially increased.

To confirm that isolated primary human gastric myofibroblasts retained characteristic morphology, markers and function of each cell line was tested to confirm the presence of a spindle/stellate morphology and strong co-expression of the two canonical myofibroblast markers, vimentin and αSMA (Figure 3.2). In addition, all CAMs were tested to ensure that they retain the ability to induce the migration and proliferation of the AGS gastric cancer cell line (Figure 3.3). Data from these studies confirm that at the time at which RNA was prepared from primary gastric myofibroblasts they maintained all of the hallmarks that have been used to define this cell type in other tissues.

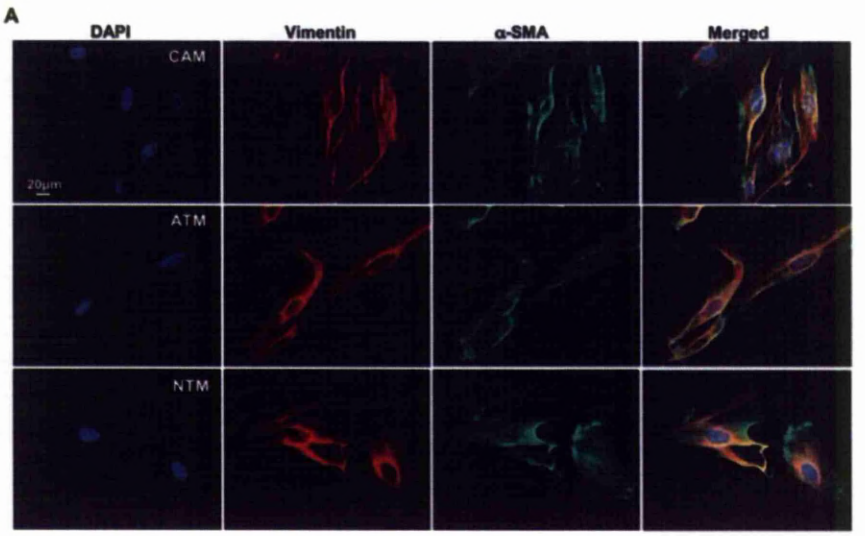



**Figure 3.2** Fluorescence images showing strong coexpression of the myofibroblast markers vimentin and α-SMA in isolated primary human CAMs, ATMs and normal tissue myofibroblasts NTMs. From Holmberg et. al. Carcinogenesis 33:1553 (2012)

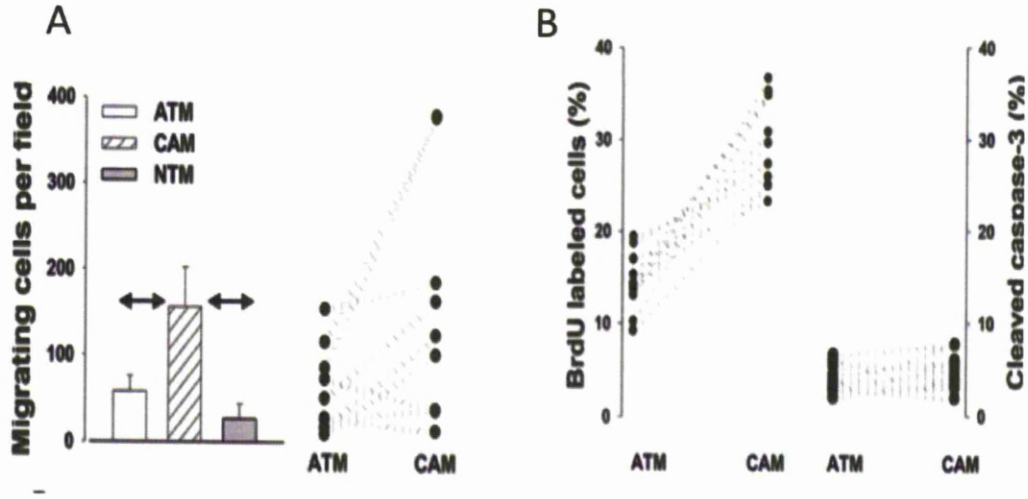



**Figure 3.3** Isolated human gastric CAM cell lines retain an ability to induce migration and proliferation of AGS gastric cancer cells. (A) Increased migration of CAMs compared with ATMs and NTMs (left) in Boyden chambers, and individual pair-wise comparisons of CAMs versus their corresponding ATMs (right). (B) Individual pair-wise comparison of BrdU labeling and cleaved caspase-3 staining in CAMs versus their corresponding ATMs. Data taken from Holberg et. al Carcinogenesis 33:1553 (2012).

Following isolation, myofibroblasts were cultured as described previously (McCaig et al.2006), before being processed for transcriptional profiling always less than passage 7. In each case cells were grown to 80% confluence before being harvested and processed for RNA isolation. Due to the number of samples being investigated myofibroblasts were processed for microarray analysis in batches as indicated in the

Table 3.1.

| CAM Label | Sample | Batch date | ATM Label | Sample | Batch date |
|---|---|---|---|---|---|
| 1-CAM | Sz42/1 P5 | 09/04/2008 | 1-ATM | Sz42/2 P5 | 04/09/2008 |
| 2-CAM | Sz45/1 P5 | 27/11/2007 | 2-ATMA | Sz45/2 P5 | 29/11/2007 |
| | | | 2-ATMB | Sz45/22 P5 | 11/07/2008 |
| 3-CAM | Sz190/1 P4 | 09/04/2008 | 3-ATM | Sz190/2 P4 | 27/11/2007 |
| 4-CAM | Sz192/1 P5 | 27/11/2007 | 4-ATM | Sz192/2 P5 | 18/12/2007 |
| 5-CAM | Sz194/1 P5 | 29/11/2007 | 5-ATM | Sz194/2 P5 | 04/09/2008 |
| 7-CAM | Sz198/1 P5 | 29/11/2007 | 7-ATM | Sz198/2 P5 | 20/02/2008 |
| 8-CAM | Sz268/1 P5 | 18/12/2007 | 8-ATMA | Sz268/2 P5 | 15/04/2008 |
| | | | 8-ATMB | Sz268/22 P5 | 15/04/2008 |
| 9-CAM | Sz271/1 P5 | 18/12/2007 | 9-ATM | Sz271/2 P5 | 15/04/2008 |
| 10-CAM | Sz294/1 P4 | 09/04/2008 | 10-ATMA | Sz294/2 P5 | 02/05/2008 |
| | | | 10-ATMB | Sz294/22 P4 | 02/05/2008 |
| 11-CAM | Sz305/1 P5 | 20/02/2008 | 11-ATMB | Sz305/22 P5 | 20/05/2008 |
| 12-CAM | Sz308/1 P5 | 09/04/2008 | 12-ATM | Sz308/22 P6 | 20/05/2008 |
| 13-CAM | Sz187/1 P8 | 20/02/2008 | | | |
| 14-CAM | Sz197/1 P5 | 20/02/2008 | | | |
| 15-CAM | Sz389/1 P7 | 15/04/2008 | 15-ATM | Sz389/2 P7 | 12/04/2009 |
| **AN Label** | **Sample** | **Batch date** | **AN Label** | **Sample** | **Batch date** |
| 21-ANMA | Sz196/2 | 09/04/2008 | | | |
| 22-ANMB | Sz241/22 P6 | 20/06/2008 | | | |
| 23-ANMA | Sz246/2 P6 | 20/06/2008 | | | |
| 23-ANMB | Sz246/22 P6 | 20/06/2008 | | | |
| 24-ANMA | Sz261/2 P6 | 20/06/2008 | | | |
| 24-ANMB | Sz261/22 P6 | 20/06/2008 | | | |
| 25-ANMA | Sz279/22 P4 | 20/06/2008 | | | |
| 26-ANMA | Sz334/2 P5 | 10/07/2008 | | | |
| 26-ANMB | Sz334/22 P5 | 10/07/2008 | | | |
| 27-ANMA | Sz351/2 P5 | 10/07/2008 | | | |
| 27-ANMB | Sz351/22 P5 | 10/07/2008 | | | |
| 28-ANM | 845 P7 | 10/07/2008 | | | |

**Table3.1** Cancer Associated Myofibroblasts (CAM), Adjacent Tumour Myofibroblast (ATM) and Absolute Normal Myofibroblasts (ANM) displaying associated Affymetrix batch dates.

## 3.3 Data analysis

### 3.3.1 Preliminary assessment of data quality

Affymetrix arrays were run and initially assessed for technical quality within the Liverpool Genome Research Centre (GRC) by Dr Lucille Rainbow. These arrays included spike-in hybridisation controls, which are used to assess the quality of the hybridisation process. Within the GRC BioB, BioC, BioD and CreX probes are used as internal controls to assess consistency and efficiency of hybridisation across Affymetrix arrays. Significantly, no array data is released from the GRC unless these quality controls are met. ANM example of the consistency of probe intensities across a selection of arrays is shown in Figure 3.4

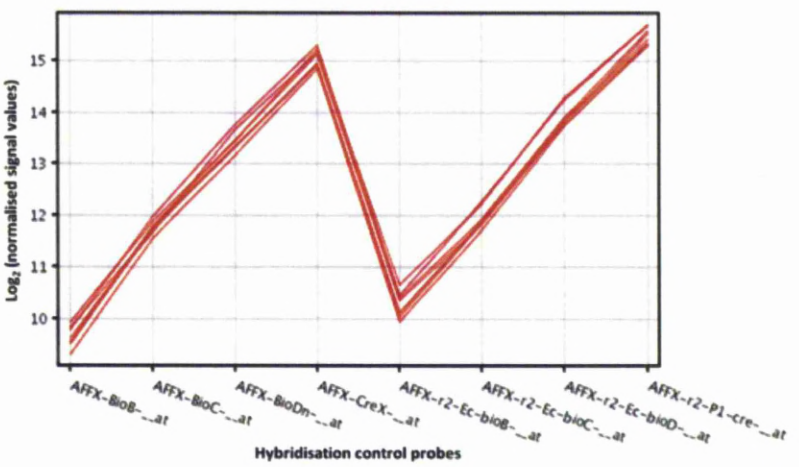



**Figure 3.4** Data from quality control tests suggest good quality hybridisation across all experiments. The normalised signal intensities of hybridisation control oligonucleotides for each microarray, shown as a red line, were comparable across all experiments. Data generated in the Varro lab (University of Liverpool).

### 3.3.2 Further detailed analysis of data quality

A range of quality control procedures are commonly used to identify, or exclude variability arising from sample preparation, processing, technical errors or extreme biological outliers. Experimental or batch variation may result from slight differences in technical procedures including RNA extraction, amplification, RNA degradation, probe-labelling, hybridization etc. In contrast, biological outliers may be indicative of contamination, incorrect sample labelling or unexpected abnormalities in the sample being studied. Determining the quality of the data is an essential first step, which is routinely performed prior to normalisation. Several methods of normalising Affymetrix microarray data have been widely used in the literature. Common examples include the RMA and Mas5 (Hubbell et al., 2002; Irizarry et al., 2003b; Pepper et al., 2007). In the RMA normalisation method data is quantile normalised (Irizarry et al., 2003c). This involves taking the average intensities of all probes across all chips, therefore a poor array would potentially skew the whole dataset, if it were not identified and removed. In contrast, in the Mas5 normalisation method each chip is independently and sequentially analysed and linear regression methods are then used to re-scale each chip to its total intensity (Hubbell et al., 2002). Although in the Mas5 method data from good quality arrays would not be directly affected by one or more poor quality arrays, normalisation without prior removal of poor quality arrays could result in artificial differential gene expression profiles.

To facilitate the identification and removal of poor quality array data, several different packages have been developed, including the open access Bioconductor

package, which utilises the R computing language to perform Array Quality Metrics (AQM) analyses. The AQM output report provides a range of different quality metrics, which together provide an assessment of the quality, intensity distributions, variance of the mean dependency, together with a range of Affymetrix specific plots (Kauffmann et al., 2009; Kauffmann and Huber, 2010), which enable the quality of data from individual arrays to be assessed and compared. To assess the quality of data from microarrays used in this study un-normalised CEL file data for all 39 samples, from cancer, adjacent or absolute normal myofibroblasts were analysed using the AQM package. The AQM package labels the arrays from 1-39 as shown in Table3.2.

| Array | Patient Label | Array | Patient Label |
|---|---|---|---|
| 1 | 3-CAM | 21 | 12-CAM |
| 2 | 4-CAM | 22 | 12-ATM |
| 3 | 4-ATM | 23 | 11-CAM |
| 4 | 7-CAM | 24 | 11-ATMB |
| 5 | 7-ATM | 25 | 2-CAM |
| 6 | 1-CAM | 26 | 22-ANMB |
| 7 | 13-CAM | 27 | 23-ANMA |
| 8 | 14-CAM | 28 | 23-ANMB |
| 9 | 1-ATM | 29 | 24-ANMA |
| 10 | 3-ATM | 30 | 24-ANMB |
| 11 | 5-CAM | 31 | 25-ANMA |
| 12 | 5-ATM | 32 | 26-ANMA |
| 13 | 9-CAM | 33 | 26-ANMB |
| 14 | 9-ATM | 34 | 27-ANMA |
| 15 | 8-CAM | 35 | 2-ATMA |
| 16 | 8-ATMA | 36 | 27-ANMB |
| 17 | 8-ATMB | 37 | 28-ANM |
| 18 | 10-CAM | 38 | 15-CAM |
| 19 | 10-ATMA | 39 | 15-ATM |
| 20 | 10-ATMB | | |

**Table 3.2.** AQM re-labelling of patient samples 1-39, to be used in combination with subsequent quality metric plots.

### 3.3.2.1 Individual array quality

A common method for visualizing the distribution of data on an array is to generate an MA plot of the data, where (M) is a ratio of the probe intensity divided by the average intensity across all arrays and (A) represents the average log intensities across arrays. The purpose of an MA plot is to investigate intensity bias, ideally we would expect the majority of intensities (plotted along the M=0 axis) to be distributed at the lower end of A. If the majority of points were distributed above or below 0, this could represent a problem with data quality. MA plots of the four best and worst arrays are displayed in Figure 3.5. Arrays were numbered from 1-39 for the AQM report, therefore array numbers represent the following patients; array 29 (24 ANA), array 26 (22 ANB), array 27 (23 ANA), array 4 (7 CAM), array 12 (5 ANA), array 18 (5 ANA), array21 (12 CAM), array6 (1 CAM). Across the worst arrays, there are slight banana shape trend across M (e.g. arrays 27 and 4), indicating the need for further normalisation. Generally, arrays with intensities above zero may represent increased background noise.

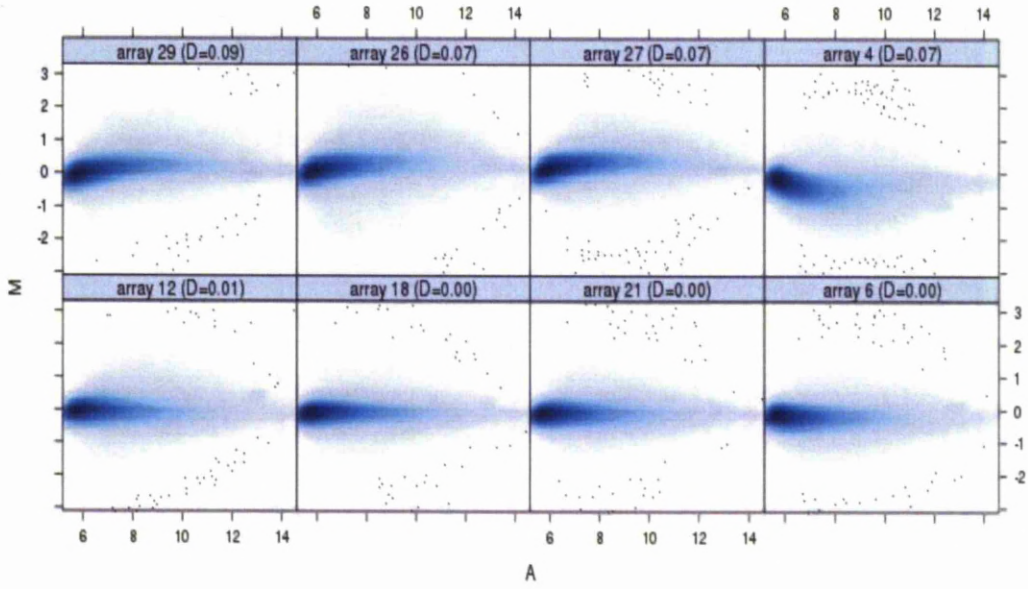



**Figure 3.5** MA plots of the four worst (upper panels) and four best plots. M (y-axis) is computed by dividing the intensity by the median intensity of the same probe across all arrays and the A (x-axis) represents the mean of their logarithms. Arrays were numbered from 1-39 for the AQM report, representing the following patients; array 29 (24 ANA), array 26(22 ANB), array 27(23 ANA), array 4(7 CAM), array 12(5 ANA), array 18(5 ANA), array21 (12 CAM), array 6 (1 CAM).

Arrays can also be visualised using spatial distributions of feature intensities (Figure 3.6). Again Spatial representations of the four best and four worst arrays are displayed, with arrays numbered from 1-39 within the AQM report, where array numbers represent the following patients; array 8 (14 CAM), array 3 (4 ATM), array 9 (1 ATM), array 2 (4 CAM), array 30 (24 ANMB), array 37 (28 ANM), array38 (15 CAM) and array35 (2 ANMA). This type of visualisation allows the easy identification of spatial effects such as air bubbles or printing problems, an obvious example of which is shown in Figure 3.7. Although array 8 (14CAM) was determined as an outlier, visually the array does not display any obvious spatial defects compared to other high-quality arrays.

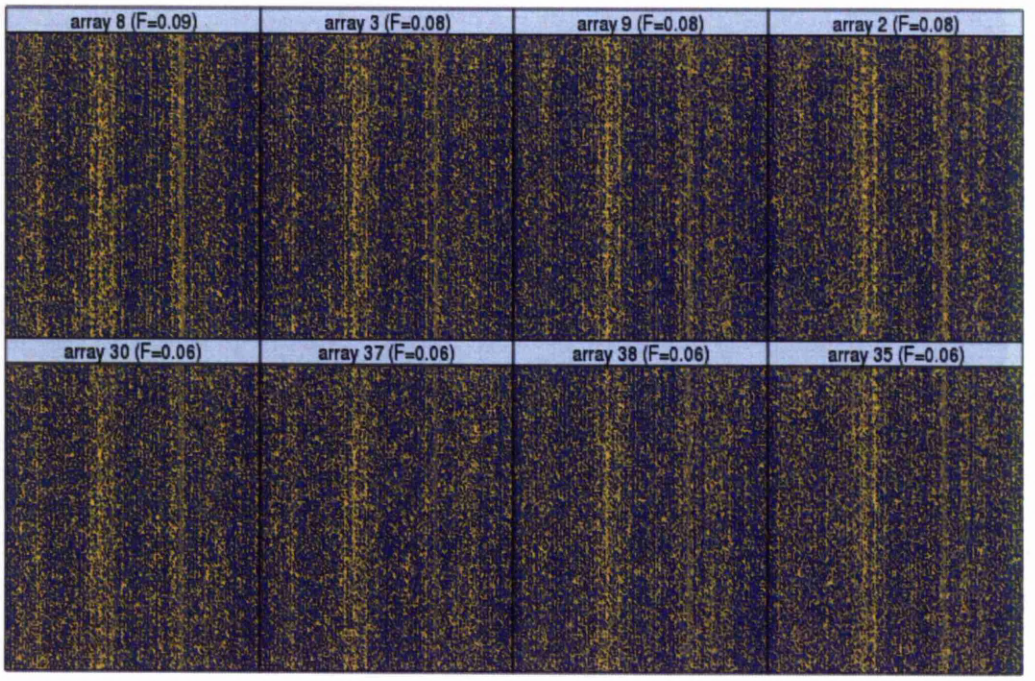



**Figure 3.6** Spatial representations of the four worst (upper panels) and four best arrays. Arrays shown above show results obtained from analysis of the following biological samples; array 8 (14 CAM), array 3 (4 ATM), array 9 (1 ATM), array 2 (4 CAM), array 30 (24 ANMB), array 37 (28 ANM), array38 (15 CAM) and array35 (2 ANA). Based on the distribution of the values across all arrays, an outlier threshold of 0.0851 was determined, which array 8 (14 CAM) exceeded.

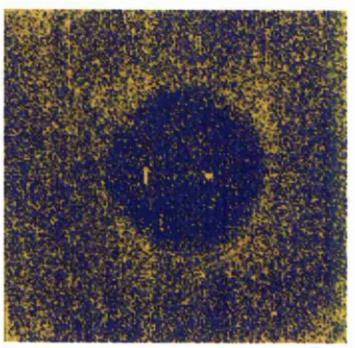


**Figure 3.7** An example of the kind of spatial defects that can occur due to the presence of an air bubble during processing. This image is not taken from an array utilized in this study.

### 3.3.2.2 Array Intensity Distributions

To assess consistency of probe intensity levels between arrays, boxplots and density plots were analysed. Boxplots for all 39 arrays are displayed in Figure 3.8A. In these plots boxes represent the interquartile range (lower quartile, median (centre line) and the upper quartile), while the whiskers represent the minimum and maximum intensities not considered to be outliers (Mcgill et al., 1978). Arrays that show consistency should have similar width and position along the x-axis, arrays with shifted positions or elongated widths may represent outliers or poor quality data. The Kolmogorov-statistic method was performed to identify statistical outliers (Figure 3.8B). Using this approach to analyses data from all arrays, a threshold of 1.65 was deemed to be the value that indicated an outlier. None of the arrays exceeded this threshold and therefore, none of the arrays analysed were considered to be statistically significant outliers.

Density plots are smoothed histograms, in which the majority of intensities are expected to be clustered around the smaller range, tailing off towards larger values. In this analysis, background noise would be seen as a shift in the distribution of the histogram towards the right, whereas small humps in the right-hand-side tail would represent saturation effects. Neither of these effects was observed in any of the plots from the 39 arrays used in this study (Figure 3.8C). Therefore, this data together with data from boxplot analysis indicate that there are no obvious outlying arrays.

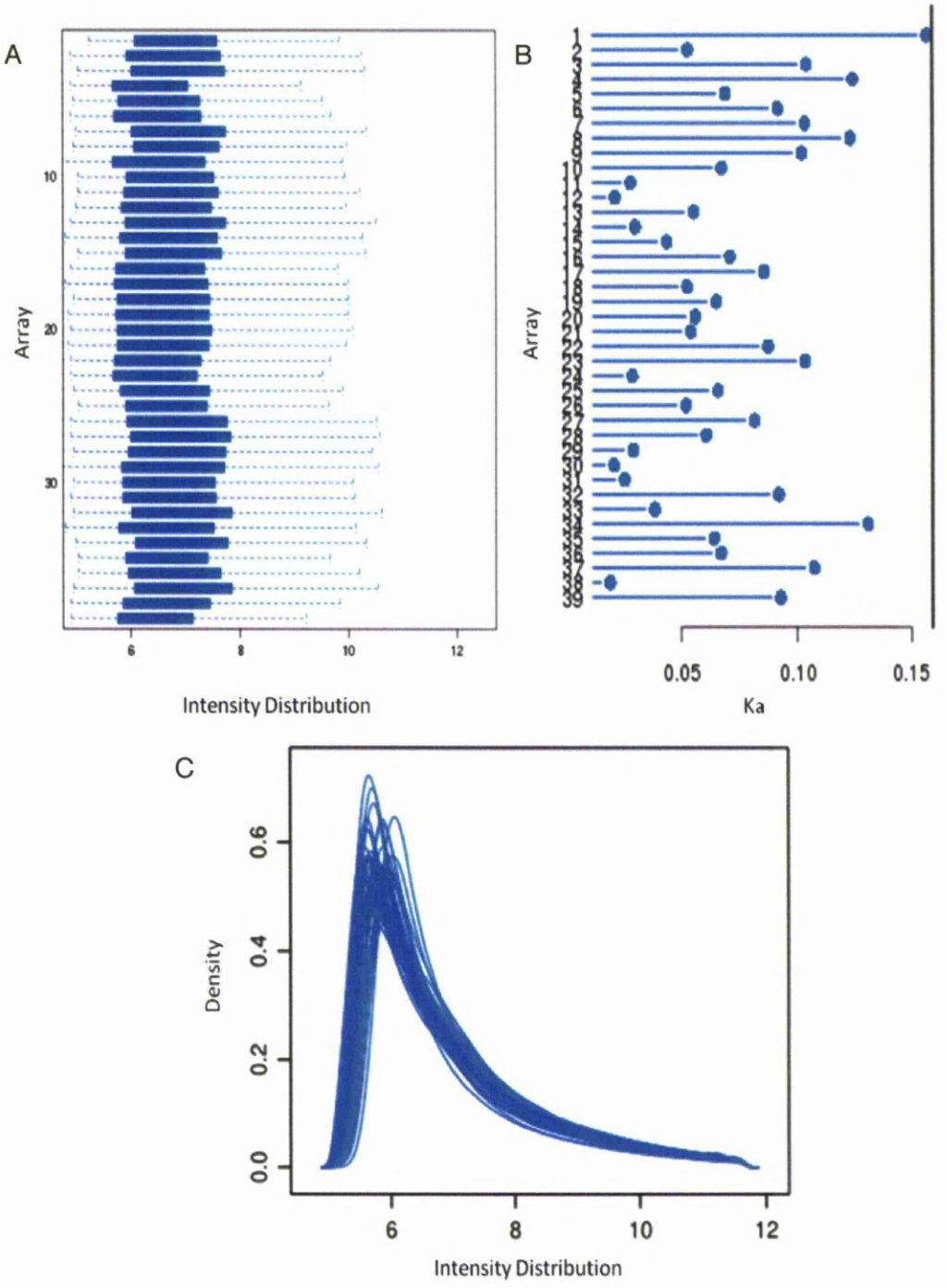



**Figure 3.8** A) Displays intensity distribution boxplots of all 39 arrays analysed in this study. B) Bar chart representing the Kolmogorov-statistic analysis for outlier detection. (Ka) Based on all the arrays, a threshold of 1.65 was deemed the value indicating an outlier. None of the arrays analysed in this study exceeded this threshold. C) Density plots (smoothed histograms) for all 39 arrays.

### 3.3.2.3 Between Array Comparisons

Hierarchical clustering and principal component analysis (PCA) methods are routinely used to identify arrays, or samples that are significantly different from one another. The Heatmap shown in Figure 3.9A represents the similarity between all 39 arrays analysed in this study, this analysis identifies two arrays (array 38 = 15 CAM and 39 = 15 ATM) as outliers. In addition, principal component analysis (Figure 3.9C) of this data identifies arrays 38 (15 CAM) and 39 (15 ATM) as being distinct from the remaining 36 arrays. These samples represent cancer and adjacent samples from the same patient (15). Although these samples have different batch dates (Table 3.1), these two microarray were carried out a year later than the other samples, therefore batch effects cannot be ruled out.

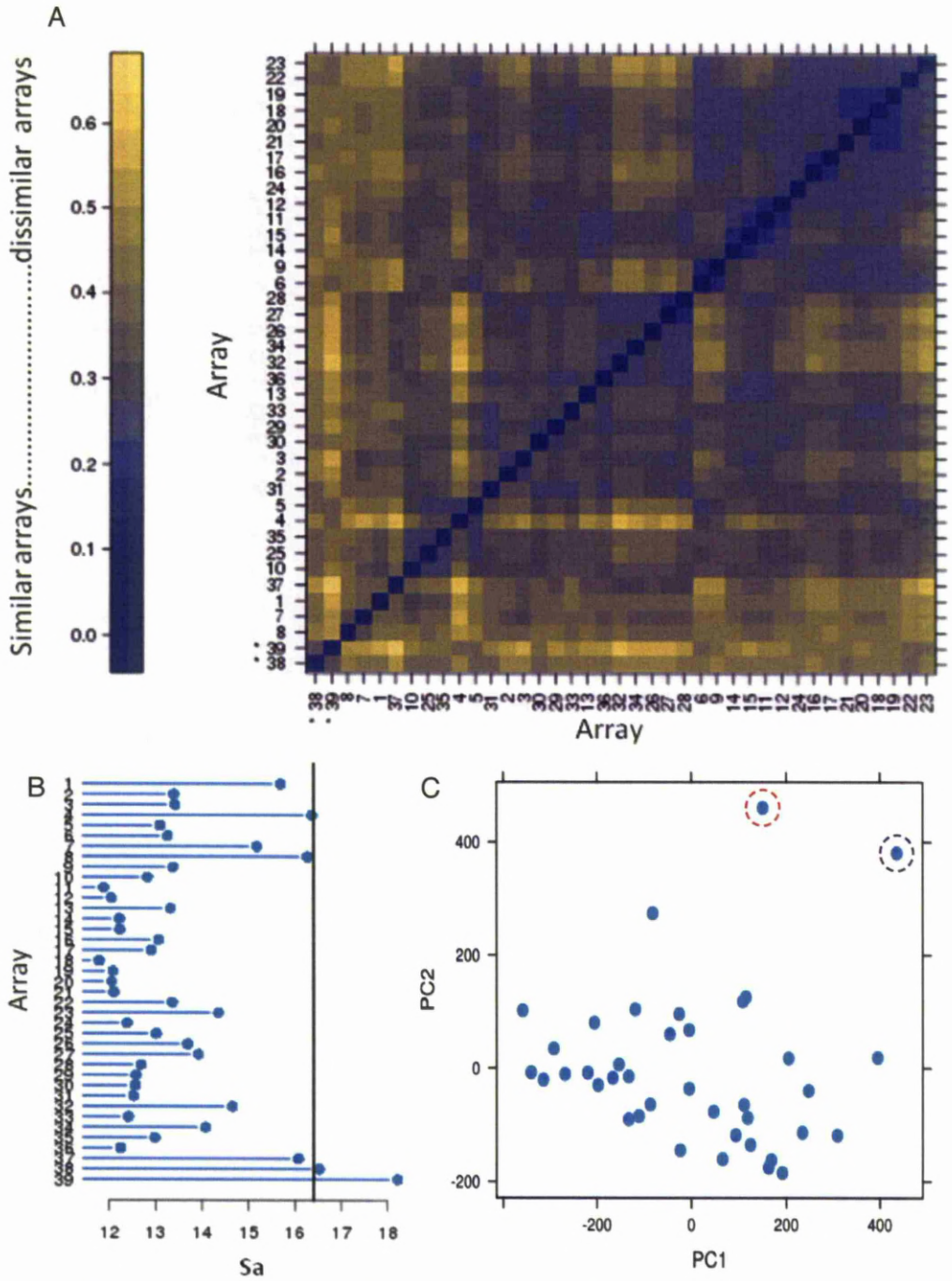



**Figure 3.9** A) A false colour heat-map showing the similarity between all 39 arrays. The colour scale on the left hand side represents the relative similarity with blue representing similar arrays whilst yellow represents least similar arrays. Outliers are identified by an asterisk. B) Bar chart representing the absolute mean distance between all arrays (Sa), computed by calculating the mean difference between every set of probes between two arrays. Arrays 38 (15 CAM) and 39 (15 ATM) were defined as outliers, as their sum of distances to all arrays is exceptionally large with bars exceeding the vertical black threshold. These samples are highlighted via an asterisk in (A). C) First two dimensions of principle component analysis visualising the similarity between arrays. Previous outlying arrays are outlined by dashed circles; array 38 (15 CAM) in red and array 39 (15 ATM) in purple.

### 3.3.2.4 Variance Mean Dependency

In variance mean dependency analyses the variance of intensities is plotted against the ranked mean of intensities (Figure 3.10). In this type of analysis, larger variations are often associated with larger intensities; therefore, a slight increasing trend in the line is expected, however the smaller the trend the greater the probability that normalisation may correct for any variability in data quality. As the trend is not strong even before normalisation, we are confident that there is no inherent bias towards intensities at either end of the spectrum in primary data used in this study.

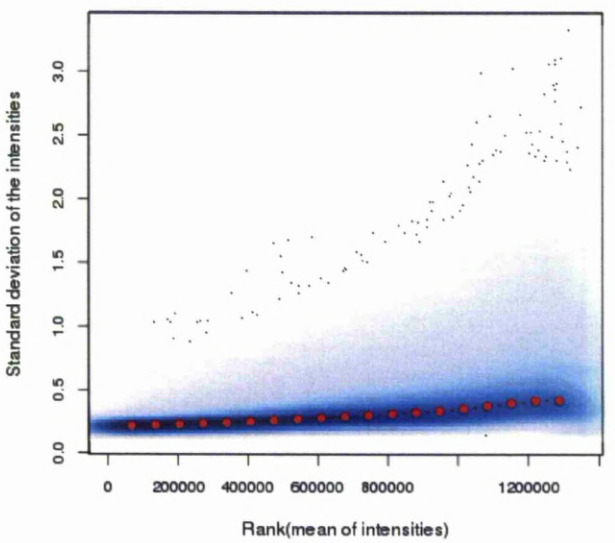



**Figure 3.10** Standard deviation for each feature on the array plotted against the mean of intensities, ranked from low to high. Red points represent the medium of the standard deviations.

### 3.3.2.5 Affymetrix Specific Plots

The final part of the AQM analysis provides a range of Affymetrix specific plots, showing relative log expression, normalised un-scaled standard error, RNA degradation and Perfect-match (PM) / Miss-match (MM) density distribution.

Logging data allows an easier comparison to be made between arrays. The relative log expression for each of the 39 arrays is displayed in Figure 3.11A. Generally, the boxes should have a similar spread and be centred around zero. However, array 8 (14 CAM), 38 (15 CAM) and 39 (15 ATM) exceeded the Kolmogorov-statistic threshold in this analysis and therefore appear to be outliers. Feature intensities as shown in Figure 3.11B also identified array 8 (14 CAM) as an outlier, although the array didn't display any obvious spatial effects. In combination with these feature intensities the variation in array intensities do not seem to be related to experimental errors. As these samples are derived from different individuals, and different gastric regions, we may expect to see a reasonable amount of biological variation between these samples. The scaled un-normalised standard errors (NUSE) for every array are shown in Figure 3.11C. The NUSE visualises the standard error of each probe sets MM>PM rate. In particular, arrays are normalised to account for the variability between chips and are adjusted so that the medians centre around one. Therefore, the greater the deviation from one the lower the array quality. In this analysis arrays 7 (13 CAM), 8 (14 CAM), 38 (13CAM) and 39 (13ATM) are defined as outliers. Therefore apart from array 7 (13 CAM), reassuringly the results obtained from the relative log expression and the normalised un-scaled standard error plots all agree with each other. Array 7(13 CAM) has performed well and has not been defined as an outlier within any of the other analytical techniques within the array quality control report. Furthermore, it is the least dramatic outlier within the NUSE plot. Therefore this sample was not excluded from subsequent analysis.

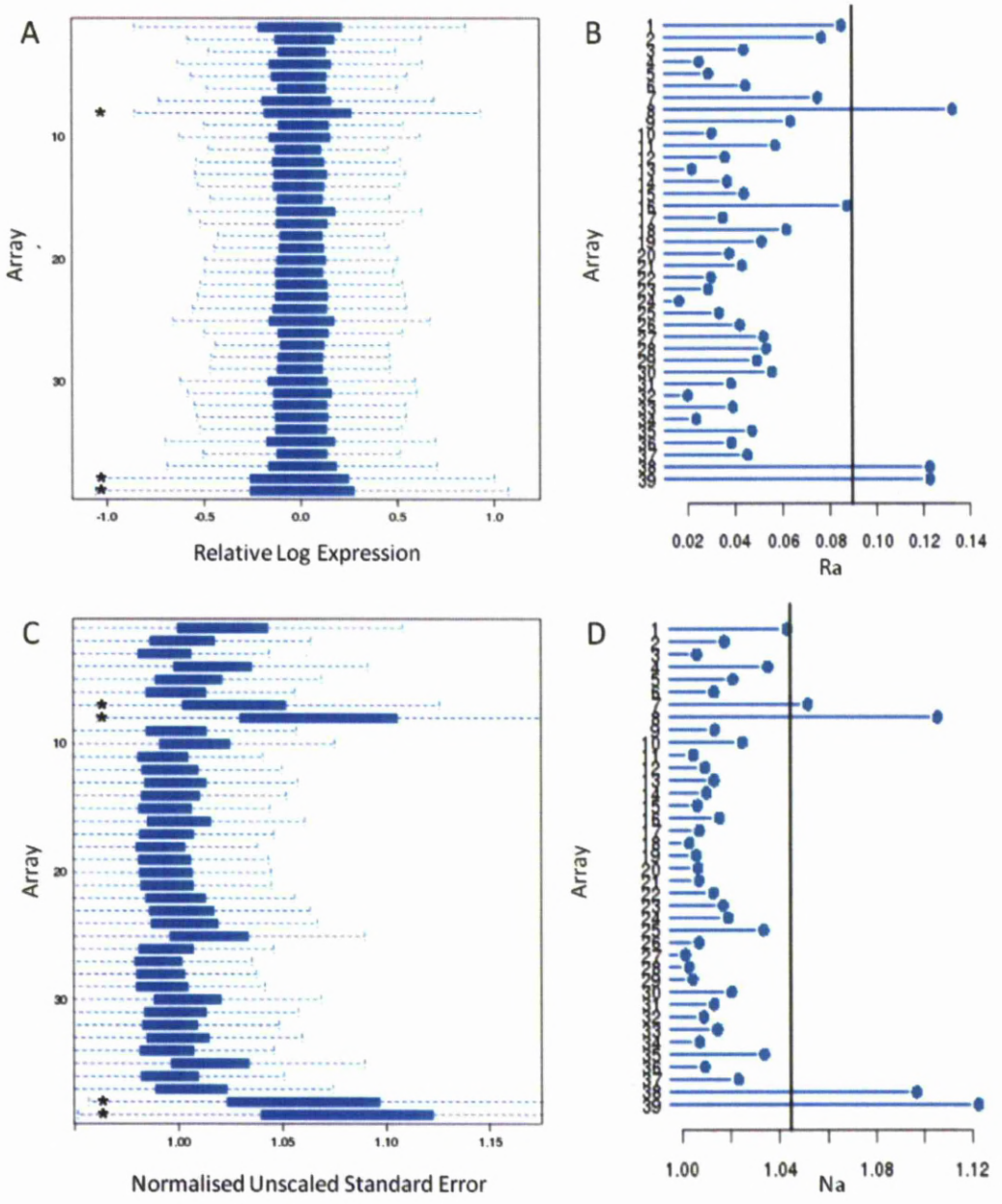



**Figure 3.11** Relative log expression for each of the 39 arrays. (A) Boxplots of individual arrays, boxes' represent the inter-quartile range, while whiskers represent the minimum and maximum intensities in each case. Arrays marked with an asterisk are considered outliers. (B) Bar chart representing the Kolmogorov-statistic for outlier detection. Based on all the arrays, a threshold of 0.0895 was deemed the value indicating an outlier. Arrays 8 (14 CAM), 38 (15 CAM) and 39 (15 ATM) exceeded this threshold and were considered outliers. (C) Normalised un-scaled standard errors represents the standard error of each probe sets MM>PM rate. Standard errors are normalised by adjusting the medians to centre around one. Arrays highlighted with an asterisk are considered outliers. (D) Outlier detection was computed using the 75$^{th}$ quantile of standard error. Based on the distribution of all of the arrays, a threshold of 1.1 was determined and arrays 8, 38 and 39 were considered outliers.

### 3.3.2.6 RNA Degradation and PM/MM Analysis

RNA degradation analysis plots the ordered probe sets by location relative to the 5' end of the target molecule against the mean intensity, based on probe location. As RNA is more often degraded from the 5' end (Dublin et al., 2000b), reduced intensities are often reported for the 5' end rather than the 3' end. This is identifiable as an increasing slope, with steeper slopes indicating a greater degree of degradation. All of the 39 arrays show encouragingly similar patterns/amounts of RNA degradation, array 38 (15CAM) and 39 (15ATM), as highlighted by the upper two lines, show slightly more, but not statistically more RNA degradation (Figure 3.12). Therefore, RNA degradation does not seem to be a significant problem in the samples used in this study.

Perfect-match (PM) and mismatch (MM) density distribution plots are shown in Figure 3.13. As expected, the density estimates of the PM intensities are shifted to the right, reflecting a greater degree of hybridisation, in comparison to MM probes, which are designed to detect non-specific binding. As the expected hybridisations have resulted from the appropriate probe type, we are confident that the probe sets are correctly detecting appropriate sequences. Together, the RNA degradation plot and the PM/MM density distribution plot display trends that are associated with good hybridisation/RNA quality.

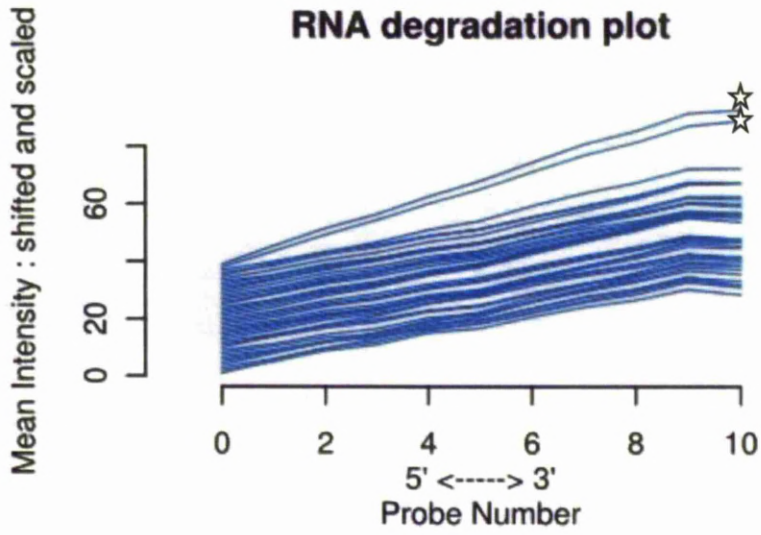



**Figure 3.12** RNA degradation plots for all 39 arrays, with each line reflecting a single chip. Probe sets are ordered by location relative to the 5' end of the target molecule (X axis) while mean intensity values of probe sets within a given location on the chip (Y axis). The two top slopes (highlighted with stars) represent array 38 (15 CAM) and 39 (15 ATM).

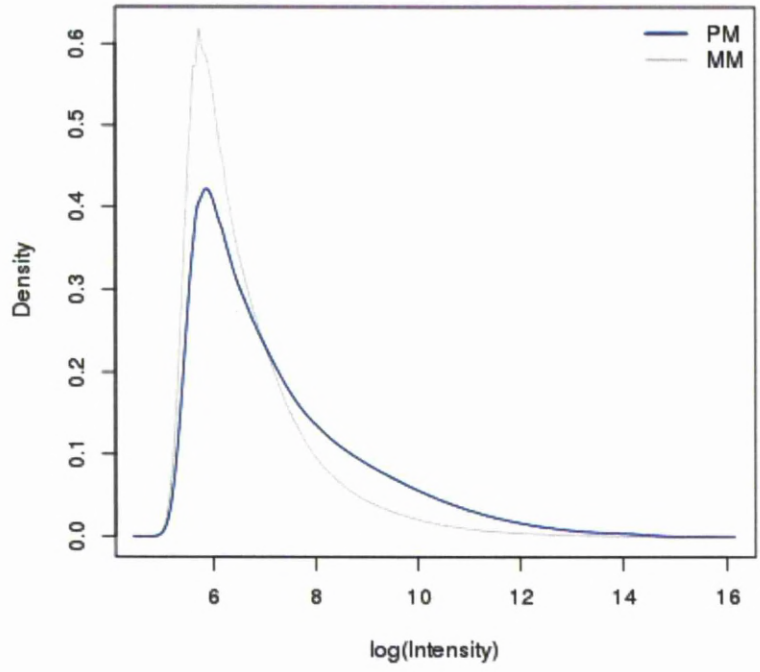



**Figure 3.13** PM/MM density distribution plot (smoothed histogram). Density estimate of intensities of PM probes are shown in blue and density estimate of intensities of MM probes are shown in grey. MM probes are expected to have poorer hybridisation than the PM probes and therefore shifted to the right.

### 3.3.3 Preliminary Principal Component Analysis

The Partek® informatics software was used to perform principal component analysis (PCA) in order to compare global gene expression signatures for all cancer (CAM), adjacent (ATM) and absolute normal (ANM) myofibroblast cell lines (Figure 3.14). In all PCAs performed in this study, cancer samples are shown in RED, adjacent myofibroblasts in BLUE and absolute normal samples in GREEN.

Before batch correction two samples did not cluster with their respective myofibroblast types, as indicated by dashed circles in Figure 3.14. Absolute normal sample Sz241/2, fell within the cancer and adjacent myofibroblast samples, whilst the adjacent sample Sz45/22 was closely associated with absolute normal samples. Following batch correction each of these samples was found to distribute with their appropriate myofibroblast groups (Figure 3.14B). However, as microarray data was generated on >3 different dates (Table 3.1), it was recommended (Dr J. Puthen, Cancer Biostatistics unit, UoL) that patient correction rather than batch correction methods would be more appropriate for this particular analysis. Regardless of the two outliers highlighted in Figure 3.14A, PCA of corrected data shows clear separation between ATM, CAM and ANMs (Figure 3.14B).

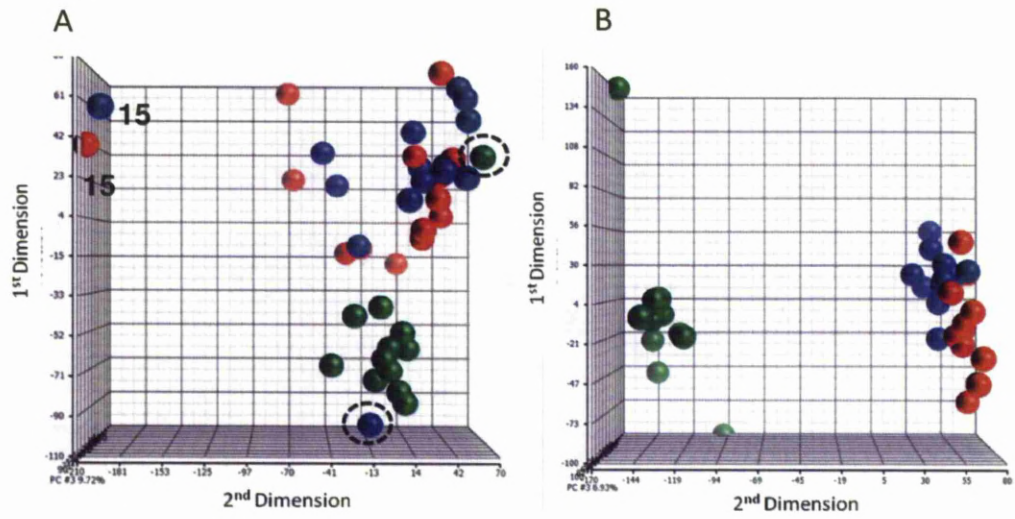



**Figure 3.14** Principal component analysis of all CAM (RED), ATM (BLUE) and ANM (GREEN) patient samples. (A) Un-corrected (B) Batch corrected. Two clear outliers are indicated by dashed circles. Also, samples labelled 15 CAM and 15 ATM were identified as outliers by Array Quality Metrics,

Samples identified as potential outliers by AQM analysis [38 (15CAM) and 39 (15 ATM)] do appear to be different from other CAM or ATM cell lines, yet both still segregate with the correct myofibroblast subgroup (samples outlined in Figure 3.14). While these differences were noted these samples were not removed as they still show distinct similarity to the correct myofibroblast populations. As global gene expression profiles for CAMs and ATM cells appeared to be very similar, a separate round of PCA was conducted on just CAM and ATM samples. Figure 5.15A, B & C show uncorrected PCA results, with cage ellipsoids showing sub-group boundaries, with an x-axis rotation being shown in Figure 3.15C. Similar to previous PCA analysis there is little distinction between CAMs and ATMs. Although, it is interesting to note that CAMs always distribute near to their associated adjacent samples (Figure 3.16A), indicating that patient specific gene expression profiles are more similar than myofibroblast sub-type expression profiles.

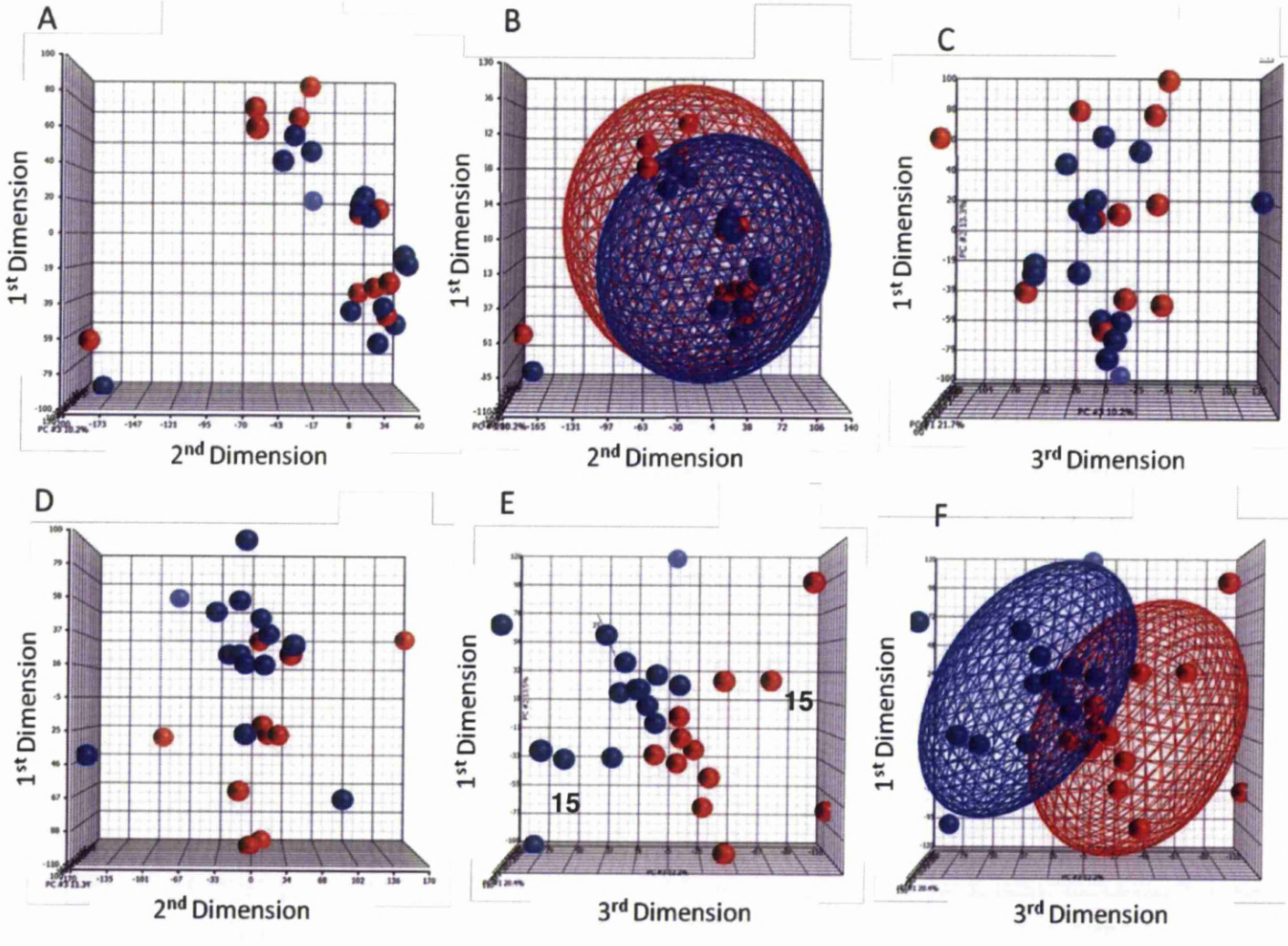



**Figure 3.15**.

Principle component analysis of CAM (RED) and ATM (BLUE) individual patient samples.

A, B and C) Show the un-corrected distribution of CAM and ATM samples, from two different angles.

D) Batch corrected CAM and ATM patient samples.

E and F) Patient Pair corrected CAM and ATM patient samples. Patient 15 CAM and 15 ATM highlighted as possible outliers within the array quality metrics report are clearly labelled.

X-axis rotation of the PCA plot shown in Figure 3.16B shows that sample Sz45/22 does not segregate close to its CAM or other ATM sample. This evidence supports the previous PCA analysis performed on all CAM, ATM and ANM samples in which sz45/22 clustered with ANM samples, rather than with other ATM samples. Therefore, ANM sample Sz241/2 and ATM Sz45/22 were removed before performing further analysis so as not to impair identification of robust subtype specific gene expression signatures.

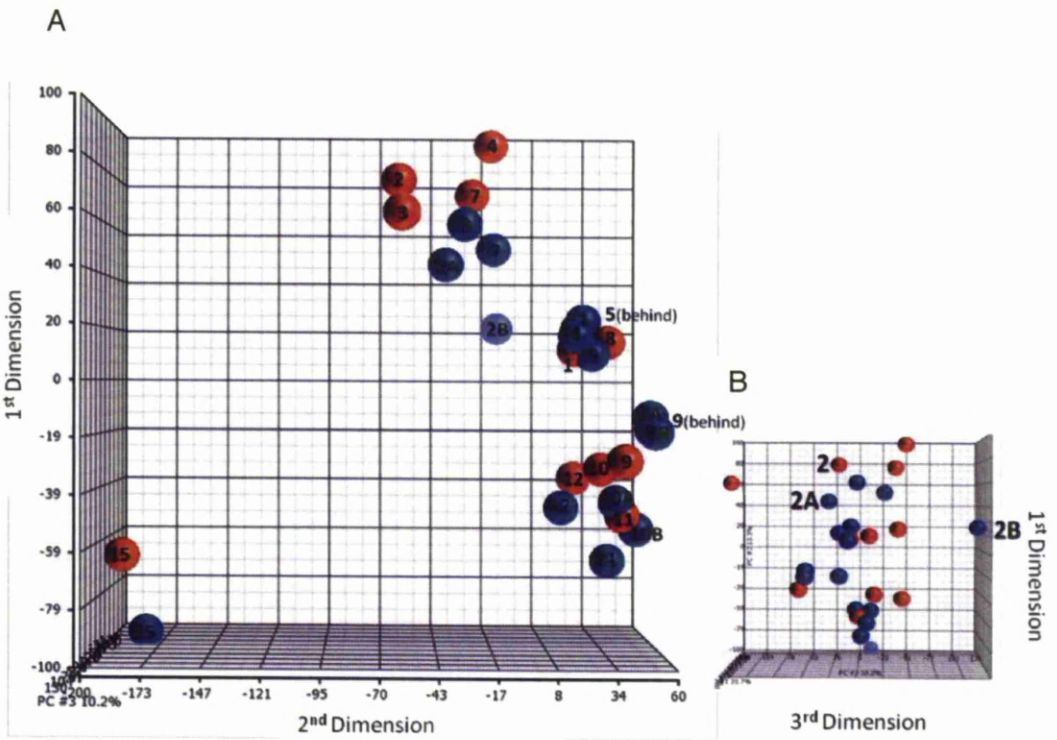



**Figure 3.16** Principal component analysis of CAM (RED) and ATM (BLUE) samples. Plots represent zoomed and annotated versions of PCA plots from Figure 5.2A. (A) Patient pairs identified (B) Rotation on the x-axis reveals position of patient label 2, 2A and 2B, relating to samples Sz45/1, Sz45/2 and Sz45/22.

### 3.3.4 Secondary Principal Component analysis

Figure 3.17 shows PCA analysis of CAM vs. ANM samples and ATM vs. ANM samples (following removal of normal sample Sz241/2 and adjacent sample Sz45/22), before and after batch correction. Without batch correction, CAM and ATM gene expression profiles are clearly distinguishable from ANM gene expression profiles. Although global gene expression profiles of CAM and ATM samples are more similar (Figure 3.18A&B), patient pairs correction improves separation of the two types of myofibroblast (Figure 3.18C&D).

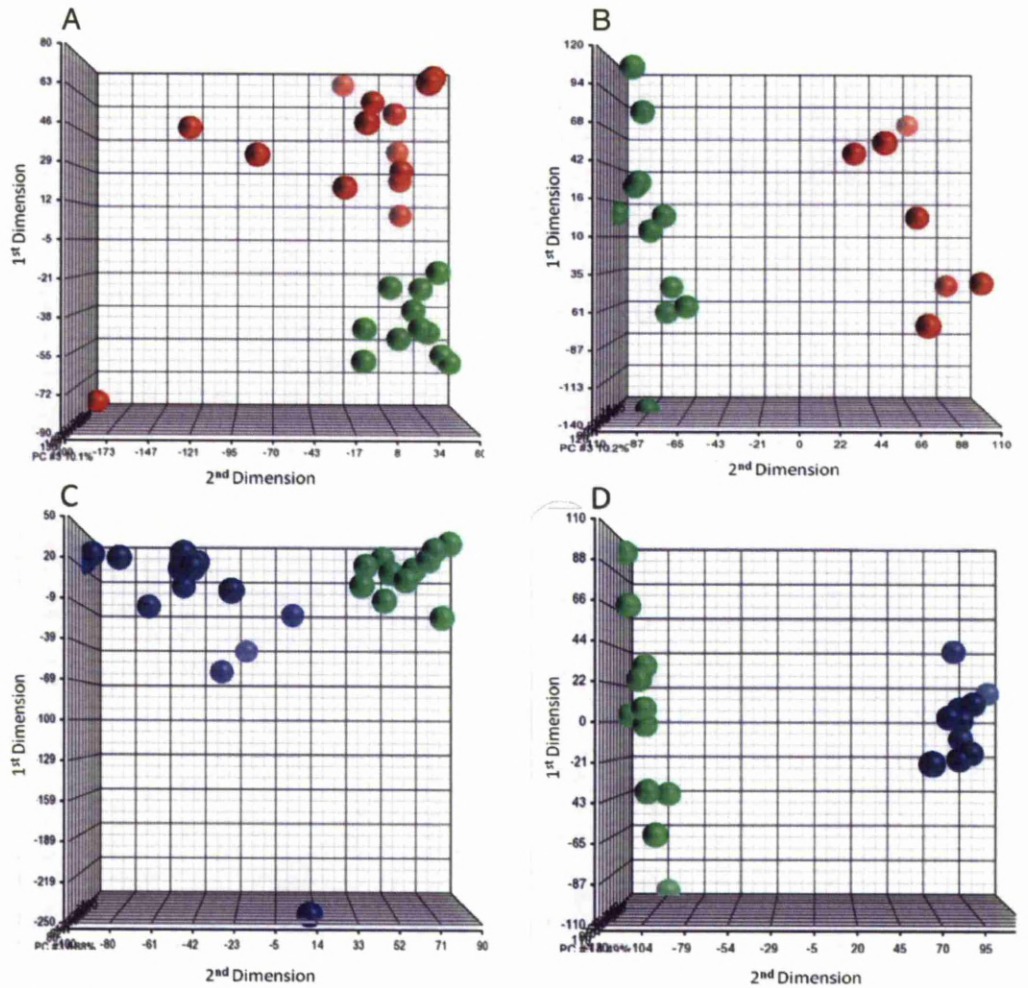



**Figure 3.17** Principal component analysis of (A) Un-corrected CAM (red) and ANM (green) samples. (B) Batch corrected CAM (red) and ANM samples (green) (C) Un-corrected ATM (blue) and ANM (green) samples (D) Batch corrected ATM (blue) and ANM (green) samples.

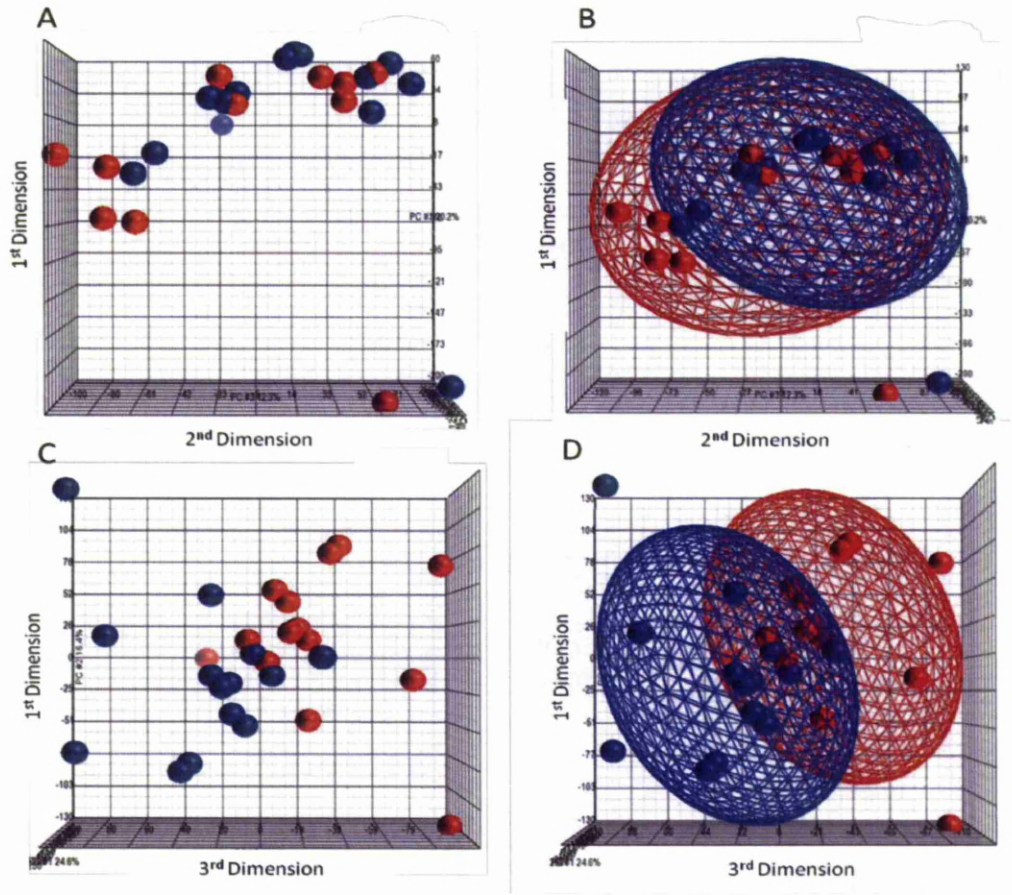



**Figure 3.18** Principal component analysis of CAM (red) and ATM (blue) samples. Panels A and B show un-corrected data while panels C and D show patient-pair corrected data. Cages represent ellipsoids, which depicts the spread of the data based on the centre of origin.

### 3.3.5 Comparison of gene expression profiles in different myofibroblast populations following Mas5 normalisation of microarray data

To compare changes occurring in CAM and ATMs, three datasets were normalised by the Mas5 method before being compiled for pairwise comparisons: CAM vs. ANM (Table 3.3A), CAM vs. ATM (Table 3.3B) and ATM vs. ANM (Table 3.3C).

A

| CAM vs. ANM | | | |
|---|---|---|---|
| Cancer | | Absolute normal | |
| Label | Sample | Label | Sample |
| 1-CAM | Sz42/1 P5 | 22-ANA | Sz241/2 P6 |
| 2-CAM | Sz45/1 P5 | 22-ANB | Sz241/22 P6 |
| 3-CAM | Sz190/1 P4 | 23-ANA | Sz246/2 P6 |
| 4-CAM | Sz192/1 P5 | 23-ANB | Sz246/22 P6 |
| 5-CAM | Sz194/1 P5 | 24-ANA | Sz261/2 P6 |
| 7-CAM | Sz198/1 P5 | 24-ANB | Sz261/22 P6 |
| 8-CAM | Sz268/1 P5 | 25-ANA | Sz279/22 P4 |
| 9-CAM | Sz271/1 P5 | 26-ANA | Sz334/2 P5 |
| 10-CAM | Sz294/1 P4 | 26-ANB | Sz334/22 P5 |
| 11-CAM | Sz305/1 P5 | 27-ANA | Sz351/2 P5 |
| 12-CAM | Sz308/1 P5 | 27-ANB | Sz351/22 P5 |
| 13-CAM | Sz187/1 P8 | 28-AN | 845 P7 |
| 14-CAM | Sz197/1 P5 | | |
| 15-CAM | Sz389/1 P7 | | |

B

| CAM vs. ATM | | | |
|---|---|---|---|
| Cancer | | Adjacent | |
| Label | Sample | Label | Sample |
| 1-CAM | Sz42/1 P5 | 1-ATM | Sz42/2 P5 |
| 2-CAM | Sz45/1 P5 | 2-ATMA | Sz45/2 P5 |
| 3-CAM | Sz190/1 P4 | 2-ATMB | Sz45/22 P7 |
| 4-CAM | Sz192/1 P5 | 3-ATM | Sz190/2 P4 |
| 5-CAM | Sz194/1 P5 | 4-ATM | Sz192/2 P5 |
| 7-CAM | Sz198/1 P5 | 5-ATM | Sz194/2 P5 |
| 8-CAM | Sz268/1 P5 | 7-ATM | Sz198/2 P5 |
| 9-CAM | Sz271/1 P5 | 8-ATMA | Sz268/2 P5 |
| 10-CAM | Sz294/1 P4 | 8-ATMB | Sz268/22 P5 |
| 11-CAM | Sz305/1 P5 | 9-ATM | Sz271/2 P5 |
| 12-CAM | Sz308/1 P5 | 10-ATMA | Sz294/2 P5 |
| 15-CAM | Sz389/1 P7 | 10-ATMB | Sz294/22 P4 |
| | | 11-ATMB | Sz305/22 P5 |
| | | 12-ATM | Sz308/22 P6 |
| | | 15-ATM | Sz389/2 P7 |

C

| ATM vs. AN | | | |
|---|---|---|---|
| Adjacent | | Absolute normal | |
| Label | Sample | Label | Sample |
| 1-ATM | Sz42/2 P5 | 22-ANMA | Sz241/2 P6 |
| 2-ATMA | Sz45/2 P5 | 22-ANMB | Sz241/22 P6 |
| 2-ATMB | Sz45/22 P7 | 23-ANMA | Sz246/2 P6 |
| 3-ATM | Sz190/2 P4 | 23-ANMB | Sz246/22 P6 |
| 4-ATM | Sz192/2 P5 | 24-ANMA | Sz261/2 P6 |
| 5-ATM | Sz194/2 P5 | 24-ANMB | Sz261/22 P6 |
| 7-ATM | Sz198/2 P5 | 25-ANMA | Sz279/22 P4 |
| 8-ATMA | Sz268/2 P5 | 26-ANMA | Sz334/2 P5 |
| 8-ATMB | Sz268/22 P5 | 26-ANMB | Sz334/22 P5 |
| 9-ATM | Sz271/2 P5 | 27-ANMA | Sz351/2 P5 |
| 10-ATMA | Sz294/2 P5 | 27-ANMB | Sz351/22 P5 |
| 10-ATMB | Sz294/22 P4 | 28-ANM | 845 P7 |
| 11-ATMB | Sz305/22 P5 | | |
| 12-ATM | Sz308/22 P6 | | |
| 15-ATM | Sz389/2 P7 | | |

**Table 3.3** (A) CAM vs. ANM dataset. (B) CAM vs. ATM dataset, Patient samples sz45/1, sz268/1 and sz294/1, were used twice for statistical analysis. (C) ATM vs. ANM dataset.

#### 3.3.5.1 Compiling background and differentially expressed gene lists

For this study background oligonucleotides were defined as all probes detected at or above baseline levels. Therefore, for oligonucleotides to be defined as present in our datasets a 'P' flag must be present in 100% of patients from either or both of the comparison groups. Numbers of oligonucleotides identified as background oligonucleotide probe IDs for each dataset are listed in Table 3.4. Student T-Tests were used to define background oligonucleotides that were differentially expressed in each comparison. For the CAM vs. ANM dataset, an unpaired student T-test was performed on 14 cancer samples vs. 12 absolute normal samples (Table 3.3A). For the ATM vs. ANM dataset, an un-paired student T-Test was performed on 15

adjacent samples and 12 absolute normal samples (Table 3.3C). For the CAM vs. ATM dataset, as patients sz45, sz268 and sz294, had a single CAM but two ATM samples, CAM sample was used twice for statistical analysis and a paired student T-Test was performed on the 15 CAM and 15 ATM samples (Table 3.3B).

| **CAM vs. ANM** | Affymetrix IDs | Entrez gene IDs |
|---|---|---|
| Background detected | 19,251 | 10667 |
| Significantly changed P-value<=(0.05) | 2958 | 2277 |
| Significantly changed P<=(0.05) and >2FC | 325 | 246 |
| **CAM vs. ATM** | Affymetrix IDs | Entrez gene Ids |
| Background detected | 18,994 | 10,723 |
| Significantly changed P-value<=(0.05) | 2364 | 1850 |
| Significantly changed P<=(0.05) and >2FC | 67 | 48 |
| **ATM vs. ANM** | Affymetrix IDs | Entrez gene IDs |
| Background detected | 19,452 | 10,765 |
| Significantly changed P-value<=(0.05) | 5357 | 3830 |
| Significantly changed P<=(0.05) and >2FC | 600 | 431 |

**Table 3.4** Background and differentially regulated gene lists for each dataset showing efficiency of conversion of Affymetrix IDs to Entrez gene IDs.

Within the CAM vs. ANM and ATM vs. ANM datasets, oligonucleotides with Benjamini-Hochberg FDR corrected p-values $\leq 0.05$ were considered to be differentially expressed. Whist for the CAM vs. ATM datasets, un-corrected p-values $\leq 0.05$ were considered to be differentially expressed. Finally, the log fold-change was calculated to determine whether differentially expressed oligonucleotides were over or under-expressed. Oligonucleotides were classified as differentially expressed if they had a p-value $\leq 0.05$. Initially no fold change cut-off was applied in order to assess the potential effect that multiple small but significant changes in

gene expression may have on pathway perturbation and pathological load within the system. Affymetrix IDs were first converted to Entrez gene IDs in order to link expression profiles to a unique gene locus.

### 3.3.6 Detection of myofibroblast markers

To validate that samples are in fact myofibroblast cell lines, the gene expression profiles were screened for the expression of the two main myofibroblast markers, $\alpha$ SMA and vimentin (VIM) (Tuxhorn et al., 2002a). As expected, both $\alpha$ SMA and VIM were expressed in all CAMs, ATMs and ANMs at a background level, as depicted by 'Present' flags for these genes in all samples. Histograms of normalised 'natural scale' expression values across all individual samples for $\alpha$ SMA and VIM are shown in Figure 3.19.

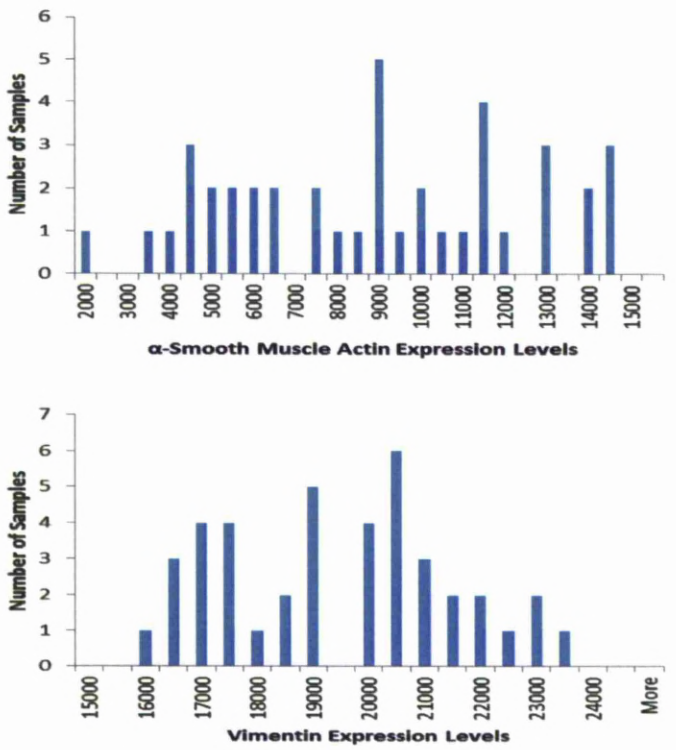



**Figure3.19** Histograms of A) alpha-smooth muscle actin (*α-SMA*) and B) Vimentin (*VIM*) Mas5 normalised 'natural scale' average probe expression values across all individual samples included in the analysis.

### 3.3.7 Results of pairwise gene expression analyses

#### 3.3.7.1 Comparing gene expression profiles between different forms of gastric myofibroblast

Following Mas5 normalisation a high degree of similarity was observed between genes that were differentially expressed in CAM vs. ANM and ATM vs. ANM comparisons. Significantly, 98.5% (1920/1948) of differentially expressed genes ($P \leq 0.05$, no fold change cut-off) were found to be changed in the same direction in both myofibroblast populations, with 901 genes being over-expressed and 1019 genes being under-expressed in both cases. In addition, 43% of genes were found to exhibit 'progressive changes' in gene expression with larger directional changes being observed in CAMs compared to ATMs. This pattern of gene expression would be consistent with a model of proximity dependant reprogramming via secreted tumour derived factors. Genes exhibiting progressive changes are discussed in more detail throughout Chapter 5.

#### 3.3.7.2 Pathway and Go annotation analysis of differentially regulated genes following Mas5 normalisation

Initially, differentially expressed genes were mapped onto pathways using the commercial Metacore™ GeneGo pathway enrichment tool. However, as approximately 70% of these genes could not be mapped to biological pathways in Metacore™ a selection of alternative bioinformatics tools including Ingenuity, Reactome, DAVID and BiNGO™ were used to provide greater insight into the range of processes that are changed in different myofibroblast populations. A comparison of the coverage provided by each tool is shown in Figure 3.20 and Table 3.5.

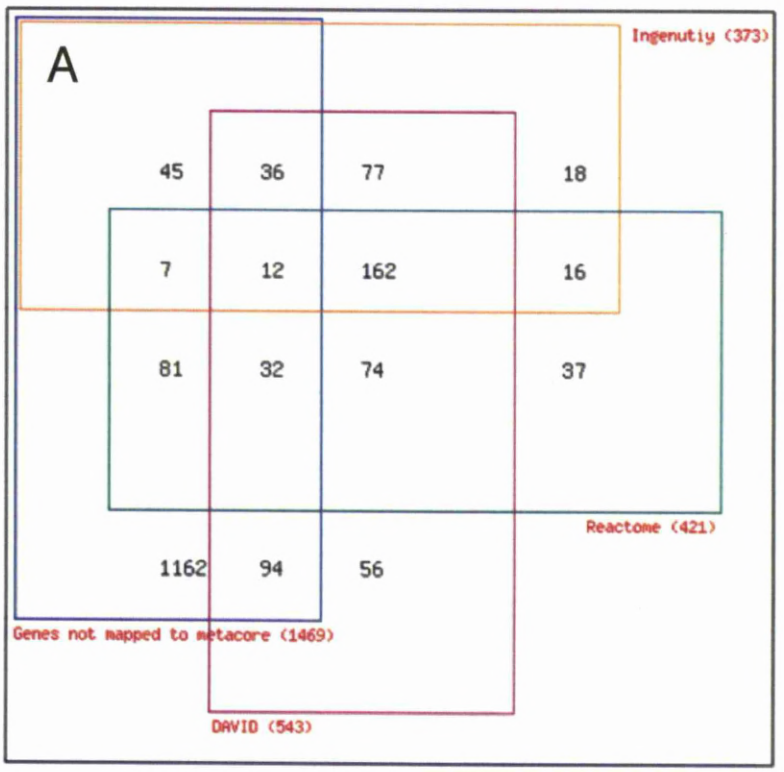



**Figure 3.20** Comparison of pathway enrichment tool coverage of CAM vs. ATM differentially regulated genes. Black numbers within rectangles represent shared differentially regulated genes. Red numbers represent total numbers of differentially regulated genes associated with each tool.

| **CAM vs. ATM** | | |
|---|---|---|
| **Metacore™ un-mapped genes** | **1469** | **% coverage** |
| DAVID | 94 | 8.506787 |
| Ingenuity | 45 | 4.072398 |
| Reactome | 81 | 7.330317 |

**Table 3.5** The numbers and percentage of CAM vs. ATM differentially regulated genes ($p<0.05$) that were un-mapped to Metacore™ canonical pathways, but are mapped to pathways within DAVID, Ingenuity or Reactome.

This process enabled a further 307 genes to be assigned to pathways/processes. This variability may be due to differences in the range of pathways available in each tool, or to differences in the compositions of pathways in different tools. Lists of significantly enriched pathways identified by each tool for the CAM vs. ATM dataset

are provided in Supplementary files 3.1-3.4, however a summary of the combined changes observed for each myofibroblast population is outlined below and key processes are discussed further in Chapter 5.

### 3.3.7.3 CAM related changes

Pathways found to be uniquely over-represented in CAMs include, TGF-β induced epithelial-mesenchyme transition (EMT). In the apoptosis and survival pathway anti-apoptotic TNFs/NF-kB/Bcl-2 and RelA transcription factors are increased, which may lead to a reduction in apoptosis via activation of BCL proteins. A decrease in integrin expression was also seen suggesting decreased cell-extracellular matrix contact. Also, three 3/4 Junction Adhesion Molecules (JAM) A, B and C were all down-regulated. This may be an important change as studies in breast cancer show that JAM A is a negative regulator of cell migration and invasion (Naik et al., 2008). Interestingly, our data also shows that cholesterol synthesis is significantly altered in CAMs. This may be important as cholesterol synthesis is up-regulated in cancer cells and the mevolanate pathway has been identified as a target for anti-cancer therapy, since Lovastatin (an inhibitor of 3-hydroxy-3-methylglutaryl-coenzyme-A reductase) was shown to reduce tumour size and has been proposed for use in phase II trials (Thibault et al., 1996).

Significantly, 3-hydroxy-3-methylglutaryl-coenzyme A reductase, already appears to be down regulated in CAMs. The potential significance of this observation is discussed further in Chapter 5. Also, 3 members of the Oligo-Saccharyl-Transferase (OST) membrane protein complex that are required for N-glycan synthesis, were

found to be up-regulated in CAMs; suggesting a role for N-glycan synthesis in gastric tumour progression.

Using the DAVID KEGG and BioCarta tools, N-glycan biosynthesis, epithelial cell signalling in Helicobacter pylori infection, leukocyte transendothelial migration, TACI and BCMA stimulation of B cell immune responses and steroid biosynthesis were found to be significantly altered in CAMs. As changes in EGF receptor glycosylation induce sensitivity to extracellular EGF (Matsumoto et al., 2008), observed changes in CAM expression profiles could explain enhanced proliferation and migration phenotypes observed in these cells. Interesting the pathway signature relating to H. pylori infection is unique to CAMs and was not detected in ATMs. Using the Ingenuity tool, NF-κB signalling was found to be altered in CAMs, with a number of genes being up-regulated, including Rel-A, which may increase inflammation, cell proliferation and survival. Clathrin-mediated endocytosis was also altered in CAMs, which may well effect growth factor signalling in these cells. The SLC16A3 and SLC7A7 monocarboxylate transporters which are responsible for the import and export of metabolites including lactate and pyruvate, which play a key role in stroma mediated tumour growth were both found to be up regulated in CAMs. The ingenuity tool also identified significant changes in the CD40 signalling pathway thus linking ER stress to increased inflammation and release of soluble factors causing the release of cytokines and subsequent T-cell activation (Hasnain et al., 2012).

A group of over-lapping pathways involved in signal transduction through the insulin receptor including PKB mediated events, mTOR signalling and AMPK activation by LKB1 all contained a large number of down regulated genes. PKB

mediated events include mTOR signalling, which conveys signals on the availability of growth factors and nutrients at the cell surface thereby controlling the rate of cell growth and proliferation. Also, an inhibitor of the mTOR complex1 raptor was up-regulated in CAMs, implying a potential decrease in mTOR mediated cell growth and proliferation in these cells.

AMPK is a key regulator of energy homeostasis, and is activated under high levels of cellular stress, or low glucose levels. AMPK inhibits glycogen and fatty acid synthesis and stimulates glycolysis and fatty acid oxidation (Steinberg and Kemp, 2009). Together these signatures provide the first insight into differential mechanisms of energy metabolism in gastric CAMs.

| Tool | Pathways/processes | Potential effects |
|---|---|---|
| Metacore™ | TGF-beta induced EMT<br>Cholesterol synthesis | Increased motility<br>Increased cholesterol |
| DAVID | *H.pylori* infection | Inflammation triggering pathway? |
| Ingenuity | P38 MAPK signalling pathway | Decreased transcription and translation |
| Reactome | HIV infection<br>Monocarboxylate transport<br>Regulation of AMPK via LKB1<br>Cytoskeletal proteins | Increased pyruvate/lactate transport<br>Inhibit storage/stimulate metabolism<br>Decreased mobility |
| Metacore™ and DAVID | N-glycan biosynthesis | Increased glycans |
| Metacore™ and Reactome | ER stress response/anti-apoptotic | Mixed apoptotic signals |
| Ingenuity and Reactome | CD40 signalling, Un-folded protein response | Increased inflammation |
| Metacore™, Reactome and Bingo | JAM adhesion molecules/ integrins/cell-cell adhesion | Decreased adhesion/increased motility? |

**Table 3.6** Summary of over-represented CAM pathways. Along with the tools used to identify specific over-represented pathways or gene ontology process and potential cellular effects.

### 3.3.7.4 Changes in gene expression observed in ATMs

Lists of significantly enriched pathways identified by each tool for the ATM vs. ANM dataset are provided in Supplementary files 3.9-3.12. Initial pathway enrichment analysis in Metacore™ revealed a significant increase in the expression of genes involved in: smooth muscle contraction, cytoskeletal remodelling and a range of metabolic pathways, including amino acid and ketone body metabolism and biosynthesis pathways which were highly up-regulated in ATMs. Together with expression of HADHB mitochondrial protein subunits, which catalyse the last steps of ketone body biosynthesis (Figure 3.21). We also observed a significant decrease in the expression of genes involved in calcium signalling pathways, and a strong over-all decrease in genes involved in DNA mismatch repair and gene expression. Significantly, Thioredoxin (TXN) and its oxidation catalyst peroxiredoxin1 (PRDX1) were up-regulated in ATMs (Figure 3.21), where they may play a key role in responding to oxidative stress within the tumour microenvironment. TXN is a redox signalling protein that can act as a messenger, in stress related signalling cascades, while also acting as an anti-oxidant reducing agent. Interestingly, the glutathione metabolism pathway was also enriched in ATMs. Glutathione acts in a similar way to thioredoxin as both play a role in orchestrating cellular responses to oxidative stress. (Lushchak, 2012). Normally cellular levels of GSH are high and GSSG are low, but under high oxidative stress this is reversed (Pompella et al., 2003). As all of the enzymes facilitating the reduction of oxidised species are up-regulated and the enzyme converting GSSG back into GSH is also increased, there is a strong indication that ATMs may be programmed to operate under conditions of oxidative stress, or exhibit unique metabolic properties to ATMs.

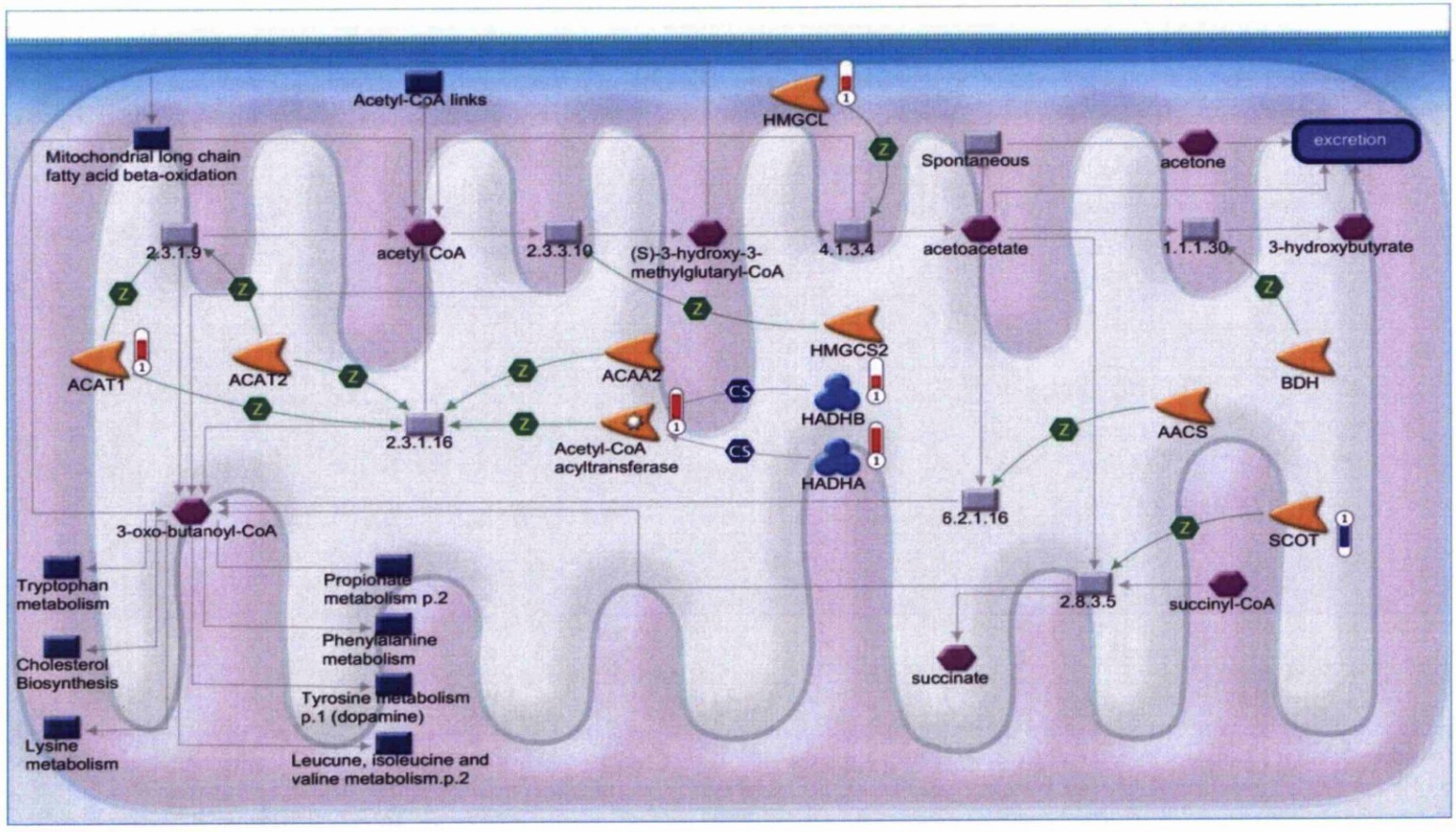



**Figure 3.21** An over-represented ($p<0.05$) ATM unique Metacore GeneGo pathway, mitochondrial ketone bodies biosynthesis and metabolism. Blue shapes represent proteins, orange represent enzymes and red represent transcription factors. Shapes labelled with a thermometer represent a differentially regulated pathway member within the CAM vs. ATM dataset. Blue thermometers represent down regulated genes whilst red thermometers represent up-regulated genes, the degree of the change is proportional to the amount of colour within the thermometer.

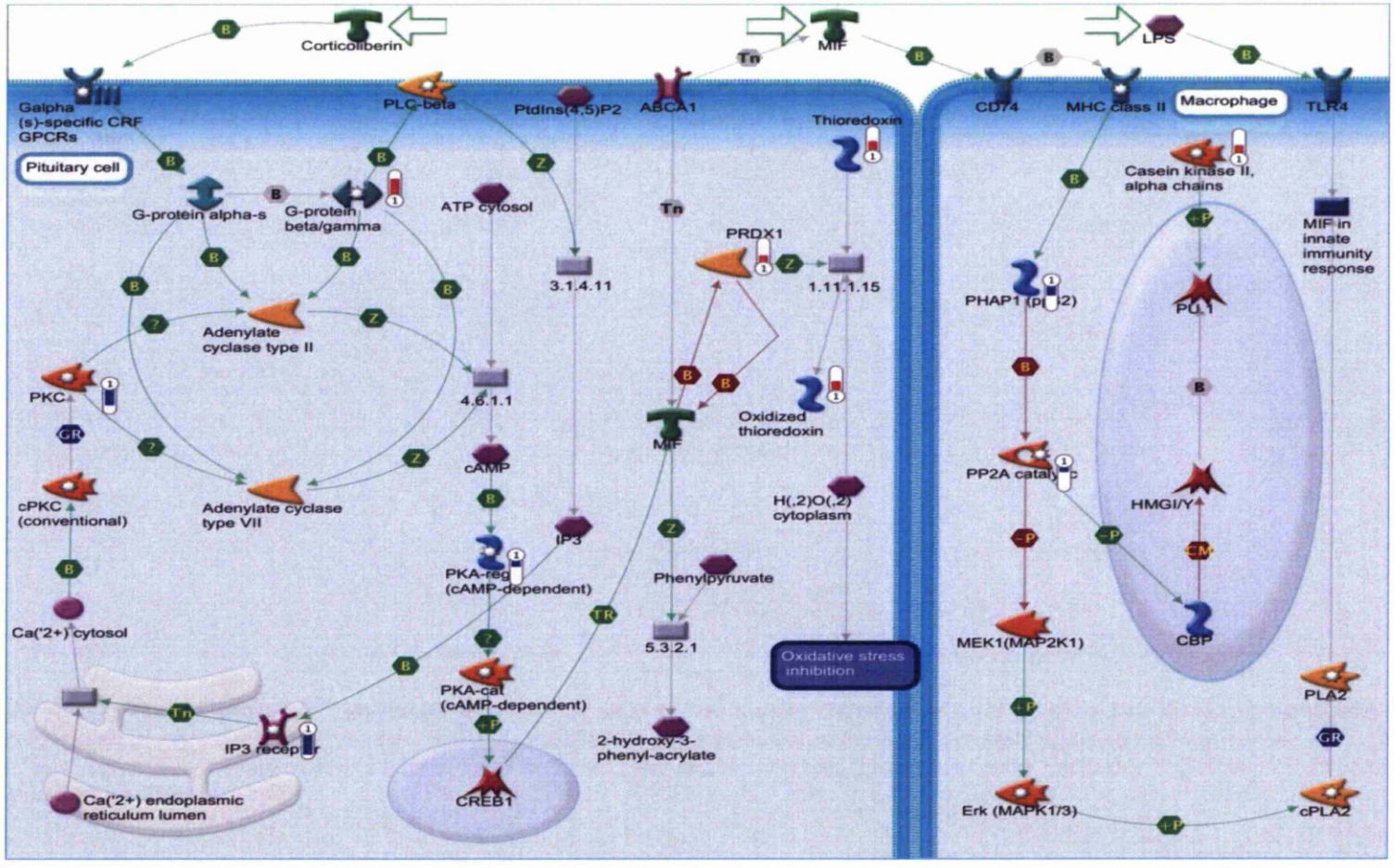



**Figure 3.22** An over-represented ($p<0.05$) ATM unique Metacore GeneGo pathway map showing the oxidative stress inhibition. Blue shapes represent proteins, orange represent enzymes and red represent transcription factors. Shapes labelled with a thermometer represent a differentially regulated pathway member within the ATM vs. ANM dataset. Blue thermometers represent down regulated genes whilst red thermometers represent up-regulated genes, the degree of the change is proportional to the amount of colour within the thermometer.

In this context it is interesting to note that Reactome analysis also identified increases in the expression of genes involved in the pentose phosphate pathway, including glucose-6-phosphate dehydrogenase (G6PD), phosphogluconate dehydrogenase (PGD) and transaldolase 1 (TALDO1). This pathway is an alternative to glycolysis, or the TCA cycle. Therefore, it is possible that ATMs have become programmed to utilise the non-oxidative pentose phosphate pathway to generate energy for themselves, whilst allowing pyruvate to be transferred to neighbouring cancer cells as proposed by the "Reverse Warburg effect". Alternatively, over-representation of the pentose phosphate shunts may be linked to levels of oxidative stress and inflammation as the end product of the pentose phosphate shunt (NADPH) also links into Glutathione metabolism.

In total 28 human solute carriers were found to be differentially regulated in ATMs. Half being up-regulated, including the pyruvate/lactate transporter SLC16A (MCT4) and half are down regulated.

Finally, downstream components of the ROCK signalling pathway are all down regulated in ATMs, suggesting that these cells have a decreased potential for migration and angiogenesis, trend supported by results recently published by Holmberg et al. (2012).

| Tool | Pathways/processes | Effects |
|---|---|---|
| Metacore™ | Smooth muscle contraction<br>DNA mis-match repair | Increased motility<br>Decreased DNA repair |
| DAVID | Primary acid biosynthesis | Increased primary bile acids |
| Ingenuity | EIF2 signalling pathway | Unclear effect on protein translation |
| Reactome | Solute carrier transport | Altered membrane transport |
| | Pentose phosphate shunt | Increased NADPH production |
| | Cell cycle | Decreased proliferation |
| | Cell junction organisation | Decreased cadherin/ cell-cell inter |
| Reactome and Ingenuity | Cellular apoptosis | Increased apoptosis |
| Ingenuity and Reactome | Thioredoxin and Glutathione | Increased oxidative stress protection |
| Metacore™ and Reactome | Ketone body metabolism | Increased Ketone bodies |
| Met, DAVID and Ingenuity | Amino acid metabolism<br>Amino acid degradation | Increased Amino acids<br>Increased Acetyl coA and ketone bodies |

**Table 3.7** Summary of over-represented unique ATM pathways, together with the tools used to identify over-represented pathway or gene ontology process and the potential cellular effects.

### 3.3.7.5 Common changes observed in both CAMs and ATMs

Complete lists of all over-represented Metacore™ GeneGo pathways, for CAM vs. ANM and ATM vs. ANM comparisons are provided in Supplementary Files 3.5 and 3.9. N-cadherin was found to be commonly down-regulated in CAMs and ATMs, along with its intracellular downstream effectors α-catenin and actin. Significantly, N-cadherin was found to be more down-regulated in CAM vs. ANM indicating a progressive change in cell-cell adhesion. A striking observation in this analysis is the fact that exactly same pathway components are frequently altered in both CAMs and ATMs, with the vast majority of changes occurring in the same direction (Figure 3.23). This potentially provides support for a model in which certain myofibroblast pathways may have been re-programmed by common insults, such as prolonged chronic inflammation, prior to the additional local reprogramming of CAMs by tumour derived factors. A comparison of the coverage provided by each tool is shown in Figure 3.24 and Table 3.8.

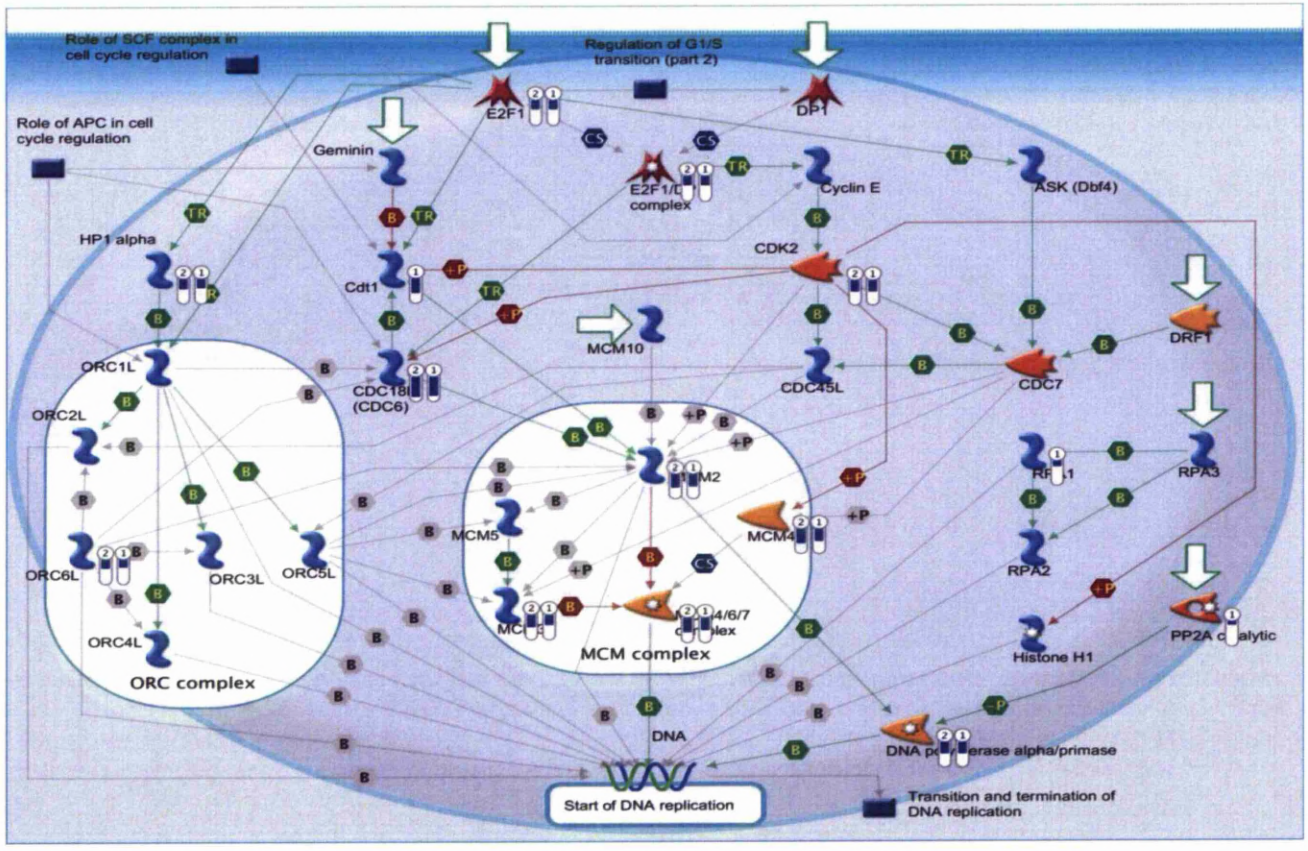



**Figure 3.23** Example of a Metacore GeneGo pathway (Cell Cycle_start of DNA replication). in which common pathway components are changed in the same way in both CAMs and ATMs. Blue shapes represent proteins, orange shapes represent enzymes and red shapes represent transcription factors. Shapes labelled with a thermometer represent a differentially regulated pathway member within the 1) ATM vs. ANM or 2) CAM vs. ANM dataset. Blue thermometers represent down regulated genes whilst red thermometers represent up-regulated genes, the degree of the change is proportional to the amount of colour within the thermometer.

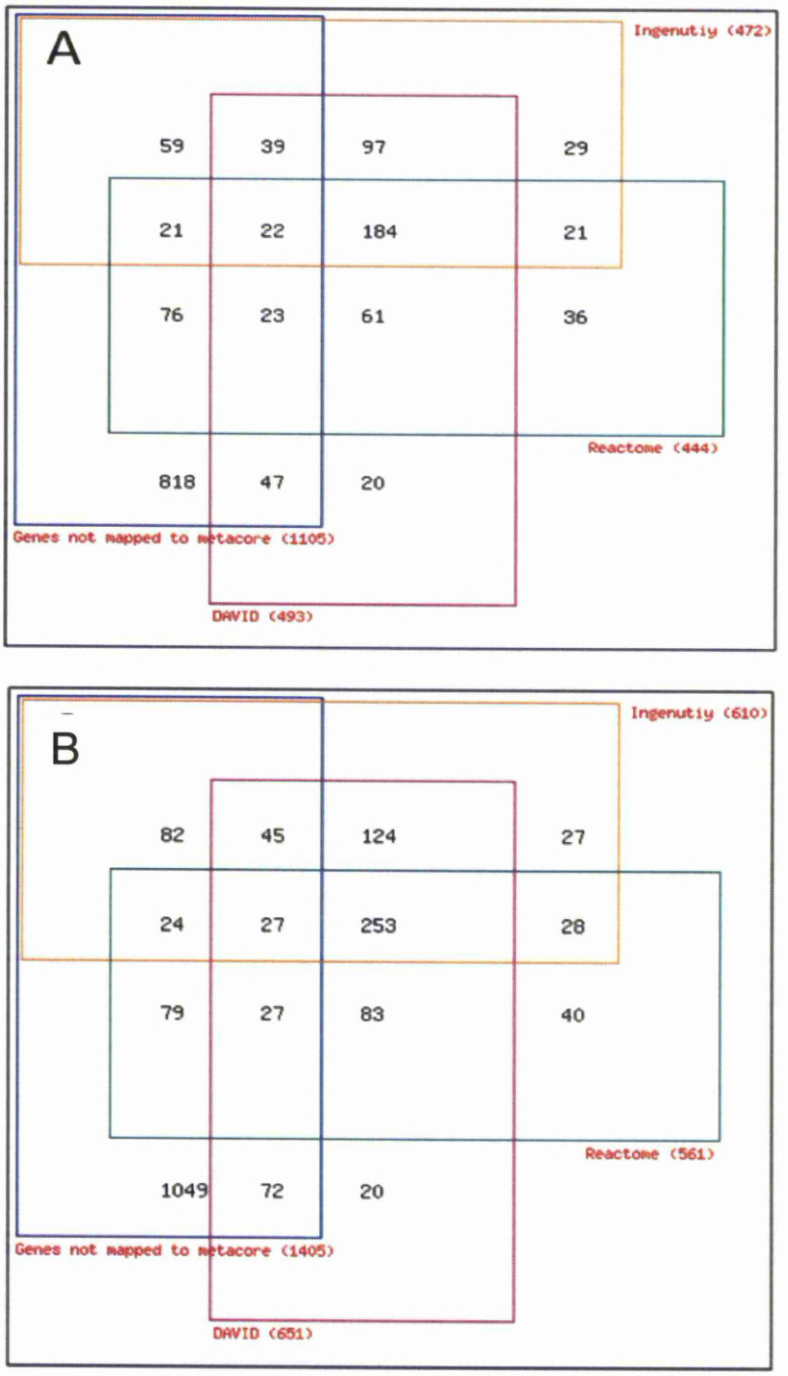



**Figure 3.24** Comparison of numbers of differentially regulated genes assigned to pathways in different informatics pathway enrichment tools (A) CAM vs. ANM. (B) ATM vs. ANM. Black numbers within rectangles represent shared differentially regulated genes and red numbers represent total numbers of differentially regulated genes associated with each individual tool.

| **CAN vs. AN** | | |
|---|---|---|
| **Metacore™ un-assigned genes** | **1105** | **% coverage** |
| DAVID | 47 | 4.253394 |
| Ingenuity | 59 | 5.339367 |
| Reactome | 76 | 6.877828 |
| **ATM vs. AN** | | |
| **Metacore™ un-mapped genes** | **1405** | **% coverage** |
| DAVID | 72 | 6.515837 |
| Ingenuity | 82 | 7.420814 |
| Reactome | 79 | 7.149321 |

**Table 3.6** Numbers and percentages of CAM vs. ANM and ATM vs. ANM differentially regulated genes ($p<0.05$), which were un-mapped to Metacore™ canonical pathways, but can be mapped to canonical pathways within DAVID, Ingenuity and Reactome.

A list of complete and statistically significant pathways ($p\leq0.05$) identified by DAVID KEGG and BioCarta, for the CAM vs. ANM and the ATM vs. ANM datasets, are provided in Supplementary Files 3.6 and 3.10 respectively. The glycosaminoglycan degradation pathway was significantly altered in both CAMs and ATMs, with several pathway components being consistently up-regulated in both cell types. Deregulation of genes involved in the glycosaminoglycan degradation pathway has previously been shown to be important in the progression of cancer (Yip et al., 2006), however the role of this process in cancer associated fibroblasts has not yet been explored. Potentially these changes could affect the biosynthesis of keratin-sulphate and heparin sulphate, which effect cell proliferation and angiogenesis (Bernfield et al., 1999), by acting as docking sites for angiogenic factors such as VEGF, FGF and MMP7, which enhance endothelial cell migration (Yu and Woessner, Jr., 2000).

Pathways found to be statistically over-represented in Ingenuity following CAM vs. ANM and ATM vs. ANM comparisons along with the complete lists are included in Supplementary File 3.7 and 3.11 respectively. Interestingly the top over-represented pathway in both the CAMs and ATMs was the synthesis and degradation of ketone bodies, with all differentially regulated genes showing an increase in expression. Ketone bodies are a by-product formed when fatty-acids are broken down to release energy, suggesting that CAMs and ATMs may both use fatty acids as an alternative form of energy to the preferred glucose. This phenomenon is discussed in more detail in Chapters 5 &6.

Pathways found to be overrepresented by Reactome analysis are included in Supplementary File 3.8 and 3.12.

| | | |
|---|---|---|
| Metacore™ | Cell-adhesion_cadherins | Decreased cell adhesion |
| DAVID | Glycosaminoglycan degradation | Increased angiogenesis |
| Ingenuity | Synthesis and degradation of ketone bodies<br>Amino acid synthesis | Increased ketone synthesis<br>Increased amino acids |
| Reactome | Influenza infection | Increased infection |
| Metacore™ and Ingenuity | DNA damage_cell cycle control | Decreased DNA repair |
| Metacore™ and Reactome | Cell-cycle | Decreased proliferation |
| Metacore™, Reactome, Ingenuity and Bingo | Translation initiation, elongation factor signalling and silencing due to infection | Increased translations, decreased if infected. |

Table 3.7: Summarisation of the common over-represented CAM and ATM pathways, along with the tools able to identify the specific over-represented pathway or gene ontology process and the potential cellular effects.

### 3.3.7.6 BiNGO™ Analysis

Even after using multiple pathway enrichment tools, a large percentage (around 65%) of genes with altered expression patterns were still unassigned to any

biological process. Therefore, BiNGO™ was used to investigate the potential function of genes within this subset of unassigned genes. For the CAM vs. ATM dataset, 678 out of 1162 differentially regulated genes were able to be assigned GO annotation terms in BiNGO™ (Supplementary File 3.13). Revealing 103 biological processes that are over-represented at the ≤0.05 level and 4 over-represented at the FDR corrected ≤ 0.05 level. Processes over-represented in the CAM vs. ATM dataset represent unique cancer changes. FDR corrected over-represented biological processes include; biological adhesion, cell adhesion, cell-cell adhesion and homeophilic cell adhesion.

For the CAM vs. ANM and the ATM vs. ANM dataset, 445 of the 818 and 564 out of 1049 differentially regulated genes were able to be assigned GO annotation terms in BiNGO™ respectively. BiNGO™ analysis of the CAM vs. ANM, revealed 169 biological processes over-represented at the ≤0.05 level and 6 over-represented at the FDR corrected ≤0.05 level (Supplementary file 3.14). BiNGO™ analysis of the ATM vs. ANM dataset, revealed 200 biological processes that were over-represented at the ≤0.05 level and 3 over-represented at the FDR corrected ≤0.05 level (Supplementary file 3.15). Processes over-represented in both the CAM vs. ANM and ATM vs. ANM datasets represent common biological processes. Very similar changes were again seen within CAMs and ATMs, with ATMs displaying a larger number of differentially regulated genes. Therefore allowing significantly more altered genes to be analysed however, findings from BiNGO™ analyses tend to agree with data obtained by all four canonical pathway tools.

### 3.3.7.7 Comparative analysis

#### 3.3.7.7.1 Genes assigned to pathways / Go annotations

The ability of each tool to map the most significantly changed genes to a pathway or process was assessed. Figure 3.25 displays the number of statistically significant genes ($p \leq 0.05$) that were mapped onto canonical pathways, within Metacore™, Ingenuity, Reactome and DAVID, or gene ontology's within BiNGO™.

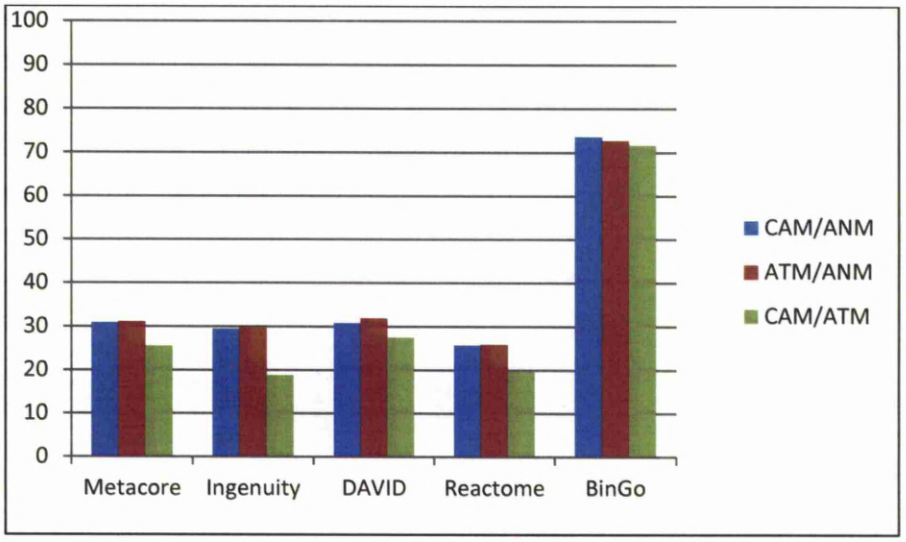



**Figure 3.25**. Percentage of differentially regulated genes ($p<0.05$), within the CAM vs. ANM (Blue), ATM vs. ANM (red) and the CAM vs. ATM (green), able to be mapped to canonical pathway analysis tools (Metacore, Ingenuity, DAVID, Reactome) and GO annotations (BinGo™).

Excluding BiNGO™, canonical pathway enrichment tools were able to map around 30% of all differentially regulated genes, across all datasets. DAVID mapped slightly more and Reactome slightly less. BiNGO™ mapped the highest number of differentially regulated genes to GO ontology terms, but as shown within BiNGO™ section 3.3.6, interpretation of the exact roles that these gene products may play within CAMs or ATMs is hard to define.

The apparent differences observed between data from different tools may be due to differences in the range of pathways available in each tools, or variability between the compositions of pathways in different tools. Overall, gene coverage within the different tools seems to lead to specific subsets of pathways being highlighted in each tool, with signalling pathways being predominantly highlighted in Ingenuity and DAVID, whilst coverage in Reactome appears to be more bias towards the cell cycle or transcription and translation and transcription related pathways.

#### 3.3.7.8 Retrospective comparison of Mas5 and RMA normalisation methods

RMA is an increasingly popular method for pre-processing array data, which includes background correction, normalisation and summarisation (Irizarry et al., 2003). There are several differences between the Mas5 and RMA methods. Mas5 normalises each array independently and uses the perfect-match (PM) and mis-match (MM) probes to assign present (P), marginal (M) and absent (A) flags alongside the normalised expressions. In contrast, RMA uses a multi-chip model, based on the theory that MM probes sometimes have higher intensity values than PM probes, resulting in negative expression values, which may cause noise at low intensity levels. Therefore, the RMA method does not consider MM probes and does not produce flagged data (Irizarry et al., 2003a).

A comparative study was performed to assess the effect that normalisation by RMA rather than Mas5 may have on the number of genes found to be significantly changed, or the range of affected biological pathways. To perform this comparison Affymetrix CEL files were uploaded into Partek® and normalised using RMA as

described in section 2.1.3.4. RMA and Mas5 normalised probes were converted into Entrez gene IDs using Metacore.

Using Mas5, only probes with a P flag in 100% of samples within a comparison group were considered for statistical analysis. In the RMA method these filters are not set, therefore all 54,000 probes on the array are considered for statistical analysis. As a result many more genes appear to be statistically changed using RMA compared to Mas5 normalisation (Supplementary files 3.16-3.18). For purposes of comparison a p-value of 0.05 and a threshold of 2-fold change were initially used to identify differentially regulated genes by both methods of normalisation.

Importantly, the CAM vs. ANM and ATM vs. ANM differentially regulated genes are defined by a Benjamini-Hochberg corrected p-value of $\leq 0.05$ and a 2 fold change in intensity, whilst the CAM vs. ATM dataset is defined by an uncorrected $p \leq 0.05$ and a 2 fold change. The Benjamini-Hochberg FDR correction cannot be applied to the CAM vs. ATM dataset as the samples are paired and too similar, therefore no genes passed an FDR corrected p-value $\leq 0.05$.

Using these criteria, the number of differentially regulated genes in each dataset, was determined using both normalisation techniques (Table 3.8). The similarity of the gene lists defined as differentially regulated using each normalisation technique are shown in Figure 3.26 Within the CAM vs. ANM dataset, 72% of Mas5 and 89% of RMA defined differentially regulated genes were identical between the two normalisation techniques. Similar percentages were also observed for the ATM vs. ANM dataset, with 72% of Mas5 and 91% of RMA defined genes being identified by both methods.

| Data Set | Mas5 | RMA |
|---|---|---|
| CAM vs. ANM | 246 | 208 |
| ATM vs. ANM | 431 | 342 |
| CAM vs. ATM | 48 | 46 |

**Table 3.8** Comparison of numbers of differentially regulated genes identified by Mas5 or RMA normalisation methods when comparing CAM vs. ANM, ATM vs. ANM or CAM vs. ATM datasets. For CAM vs. ANM and ATM vs. ANM differentially regulated genes were defined using a Benjamini-Hochberg corrected p≤0.05 and 2 fold change, whilst the CAM vs. ATM dataset is defined using an uncorrected p≤0.05 and a 2 fold change.

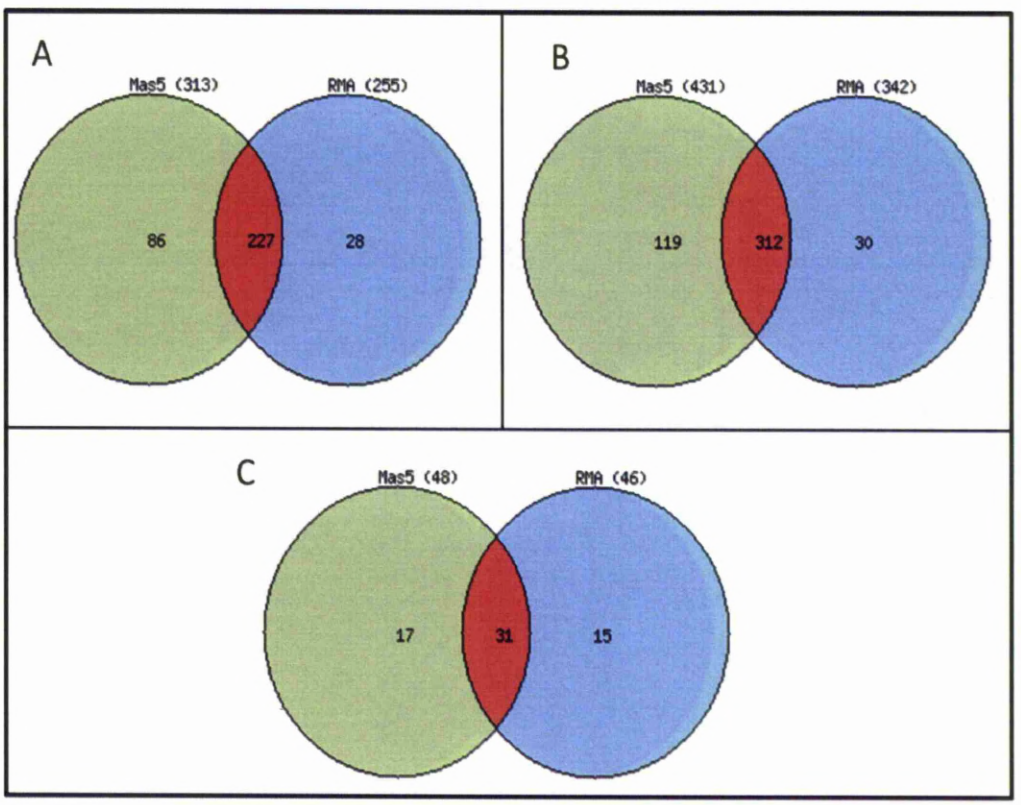



**Figure 3.26** Similarity between differentially regulated gene lists derived from Mas5 and RMA normalisation. (A) CAM vs. ANM, (B) ATM vs. ANM, (C) CAM vs. ATM. The CAM vs. ANM and ATM vs. ANM differentially regulated genes are defined using a Benjamini-Hochberg corrected p≤0.05 and 2 fold change, whilst the CAM vs. ATM dataset is defined by an uncorrected p≤0.05 and a 2 fold change. Numbers within the circles represent genes defined as uniquely differentially regulated or commonly differentially regulated following Mas5 or RMA normalisation. Numbers in brackets indicate total numbers of differentially regulated genes identified by each normalisation algorithm.

When using these stringent thresholds (p-value≤0.05 and 2 fold change), Mas5 identifies a greater number of changed genes, and RMA generated a smaller percentage of unique changed genes.

To investigate the effect that the observed 20-30% differences in changed genes may have on the range of biological processes that may be affected, the top over-represented biological pathways from a Metacore GeneGo pathway analysis were compared for gene lists generated by both normalisation methods, in each case using the entire complement of Affymetrix chip genes as a background gene list to identify biological pathways associated with the differentially regulated gene lists from both normalisation techniques.

Results from this analysis show that, although only 50% of pathways defined as statistically over-represented ($p \leq 0.05$) are identical when comparing Mas5 to RMA (CAM vs. ANM and ATM vs. ANM datasets), similar biological processes are over-represented in each case (Supplementary files 3.19-3.21). For example, in the CAM vs. ANM dataset, both techniques identified regulation of the cell cycle, DNA damage repair pathways, specifically relating to the cancer related *BRCA* genes, immune response, and VEGF signalling as top ranking changed processes. The only process uniquely identified by RMA generated data was hormone biosynthesis. With respect to the ATM vs. ANM dataset, both normalisation techniques again identified similar biological processes such as control of cell cycle, DNA damage repair, extracellular-matrix remodelling, transcriptional control of cholesterol biosynthesis, cholesterol, hormone and bile acid biosynthesis. However, in this analysis RMA normalised data uniquely identified EMT transition as a significantly changed process. Clearly this is a highly relevant process in terms of the cancer microenvironment, which would have been missed through the application of Mas5 normalisation, under these stringent conditions.

Analysis of CAM vs. ATM data normalised by Mas5 or RMA produced very similar pathway profiles, with 5/6 pathways being defined as over-represented by both normalisation techniques including developmental pathways involving epithelial mesenchymal transition and bile acid biosynthesis. The only pathway that was uniquely identified by either normalisation technique was the WNT signalling pathway identified by the Mas5 method and an immune response pathway that was only identified by RMA normalised data.

Overall, strikingly similar biologically processes were identified for Mas5 or RMA derived gene lists, despite the observed 20-30% difference in numbers of changed genes. This suggests that the 20-30% of genes that are different must either be involved in similar processes or be highly dispersed among many different pathways in such a way that these changes would not be highlighted by statistical pathway enrichment methods. Therefore, at this level of stringency, Mas5 and RMA normalisation methods result in different numbers of significantly changed genes, however (at least with respect to our data), the vast majority of biological processes found to be significantly affected is very similar. As the RMA method is not based on the use of flags to identify genes present in all samples, this method inevitably leads to the identification of significantly larger numbers of changed genes. As such, it is possible that analysis of less stringently filtered data from both normalisation methods would reveal far greater differences in both numbers of changed genes identified and the spectrum of biological processes that are changed. Data presented in Table 3.9 shows that this is the case. Using a p-value≤0.05 with no fold change cut off, the similarity of gene lists identified using Mas5 and RMA are shown in Figure3.27. Within the CAM vs. ANM and ATM vs. ANM dataset, similar

percentages of genes were defined as identically differentially regulated, around 74% and 81% respectively. However, a smaller percentage of identically differentially regulated genes (55%) were observed for the CAM vs. ATM dataset.

| Data Set | Mas5 | RMA |
|---|---|---|
| CAM vs. ANM | 2277 | 2563 |
| ATM vs. ANM | 3830 | 4264 |
| CAM vs. ATM | 1850 | 2531 |

**Table 3.9** Comparison of the number of differentially regulated genes identified by Mas5 or RMA normalisation methods when comparing CAM vs. ANM, ATM vs. ANM or CAM vs. ATM datasets. For CAM vs. ANM and ATM vs. ANM differentially regulated genes were defined using a Benjamini-Hochberg corrected p≤0.05, whilst the CAM vs. ATM dataset is defined using an uncorrected p≤0.05.

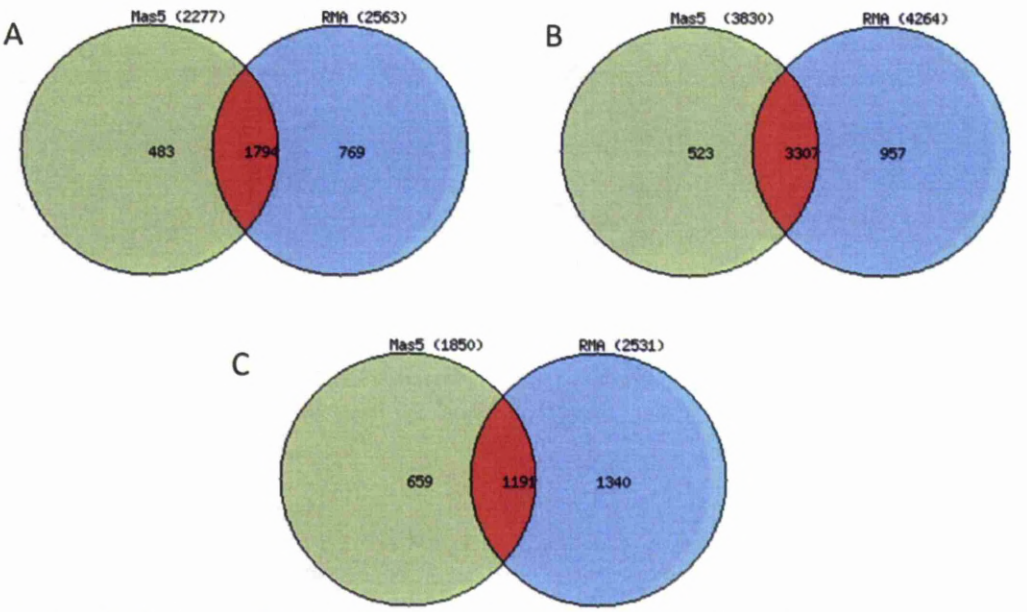



**Figure 3.27** Similarity between differentially regulated gene lists derived from Mas5 and RMA normalisation. (A) CAM vs. ANM (B) ATM vs. ANM (C) CAM vs. ATM. The CAM vs. ANM and ATM vs. ANM differentially regulated genes are defined using a Benjamini-Hochberg corrected p≤0.05, whilst the CAM vs. ATM dataset is defined by an uncorrected p≤0.05. Numbers within the circles represent genes defined as uniquely differentially regulated or commonly differentially regulated following Mas5 or RMA normalisation. Numbers in brackets indicate total numbers of differentially regulated genes identified by each normalisation algorithm.

Comparative pathway enrichment analysis of these differential genes lists in comparison to the p-value≤0.05 and 2FC, within 2/3 datasets lead to more pathways being identified within RMA than Mas5 (Supplementary files 3.22-3.24).

Upon reducing the stringency to just a p-value threshold, within the CAM vs. ANM dataset both normalisation techniques revealed the over-representation of fatty acid metabolism pathways, which we propose are important for the CAM phenotype (Chapter 5). Interestingly data from the RMA normalisation method uniquely identified K-RAS signalling and the IGF signalling pathways, both know to play important roles in cancer progression. With respect to the ATM vs. ANM dataset, both normalisation techniques revealed over-representation of a large number of lipid biosynthesis pathways and related intracellular cholesterol transport pathways, apoptosis related signalling pathways, TGF and insulin-like growth factor signalling as being significantly perturbed. The differentially regulated gene list derived from RMA normalisation revealed a number of important cancer related signalling pathways, including vascular endothelial growth factor, hedgehog, epithelial growth factor and wnt signalling.

The less stringent RMA and Mas5 derived CAM vs. ATM dataset both revealed a large set of TGF induced EMT pathways specifically in relation to cancer cells, apoptosis signalling, cell adhesion, cholesterol biosynthesis and cholesterol transport. However, RMA again uniquely revealed some biologically relevant pathways including; platelet derived growth factors in cell migration, N-Glycan biosynthesis and transforming growth factor induced cell proliferation.

Subsequent experimental work in the Sanderson lab has confirmed trends and conclusions derived from our Mas5 normalised data (Chapters 5 and 6), further experiments will now be performed to investigate potentially interesting new leads that have emerged from this retrospective reanalysis of data by the RMA method.

## 3.4 Discussion

The aim of the work described in this chapter was to establish if isolated primary human CAMs and ATMs exhibit distinct gene expression profiles, which may provide insights into the range of biological processes that are altered in comparison to gastric myofibroblasts isolated form histologicaly normal tissue. To achieve this aim several important criteria had to be satisfied, in order to be confident in predictions arising from the comparative analysis of complex gene expression profiles. Firstly, it was important to establish that isolated CAMs and ATMs retained characteristic markers and properties, which have been described for CAMs, ATMs or ANs in other tissues. Analysis of comparable low passage (p4-p7) CAM, ATM and ANM performed in the Varro lab (University of Liverpool) showed that all isolated primary human myofibroblasts retained strong co-expression of the classic myofibroblast markers vimentin and a-SMA (Figure 3.2). In addition, functional analysis of each myofibroblast cell line showed that all isolated CAMs retained the ability to enhance the migration and proliferation of AGS gastric cancer cells (Figure 3.3). Analysis of data from microarray studies also confirmed that both vimentin and a-SMA expression was detected in all of the isolated primary myofibroblast samples (Figure 3.19). Therefore, this data demonstrates that at the time that samples were prepared for microarray analysis, isolated primary human gastric myofibroblasts exhibited appropriate myofibroblast morphology, markers and functional properties.

Other criteria that needed to be addressed were the quality of the primary array data and the degree of variability that may arise from batch processing. In a retrospective analysis we also assessed the relative merits of using RMA rather than

Mas5 to normalise microarray data. As all arrays were run by the Liverpool Genome Research Facility spike-in hybridisation controls were automatically run to assess consistency across all experiments. However, the Bioconductor 'ArrayQualityMetrics' (AQM) package was also used to assess the consistency and quality of data across all 39 CAM, ATM and ANM un-normalised microarray CEL.files. This analysis showed that, Array 8 (14 CAM), 38 (15 CAM), and 39 (15 ATM) were detected as outliers due to observed variations in feature intensities (array 8 only), the relative log expression values and the observed degree of scaled un-normalised standard error. The combinations of these observations suggest that these arrays should possibly have been removed from subsequent analysis to avoid skewing results. However, it is significant to note that these two arrays were not derived from the same batches (Table 3.1). Also, following normalisation of data by either batch or patient correction methods these samples correctly segregated with appropriate myofibroblast subsets (CAMs & ATMs respectively) and neither sample was an outlier within its subtype cluster.

Considering the results of this analysis, overall the quality of the array data appears satisfactory, with the exception of 2 arrays. In retrospect it would have been preferable for samples to be processed in fewer batches, thereby making batch correction simpler and potentially more efficient. However, despite the large numbers of batch processing dates normalisation and batch/patient correction does largely correct for observed outliers in primary data. Therefore, although fewer batches would have led to less variation between samples within each myofibroblast subgroup (CAM, ATM or ANM) differences observed in gene

expression between myofibroblast subgroups appear valid, as PCA analysis shows segregation of CAM, ATM and ANM samples.

It is likely that fewer batch numbers may have reduced observed variability between samples within each myofibroblast subgroup, thereby improving our ability to identify gene expression profiles that correlate with prognostic trends, (see Chapter 5). However, in this context it is important to note that there is no obvious correlation between batch processing dates and our ability to segregate prognosis subgroups, based on differential trends in gene expression profiles. This point is discussed in more detail in Chapters 5 and 6.

Considering the consequences of using Mas5 or RMA normalisation methods. Using stringent thresholds for data from each normalisation method (FDR corrected pValue≤0.05 and 2 fold change) generated relatively similar gene lists which when used for pathway enrichment analysis also identified very similar lists of significantly affected process. However, comparison of post normalisation gene lists generated using a FDR corrected pValue≤0.05 only shows that a far greater number of altered genes were retained after normalisation. In this case RMA normalised data also identified a greater range of changed pathways, which were not identified by Mas5 normalised data. Given the larger number of genes retained following RMA normalisation it is highly likely that this dataset would also be more valuable that Mas5 data when performing correlation and correspondence studies described in chapters 4 and 5. Although Mas5 normalisation generates significantly smaller post normalisation gene lists it is likely that trends or signatures are correct, although

quite possibly incomplete. This possibility will be explored as part of on-going studies in the Sanderson lab.

In this study a range of pathway enrichment tools were used to provide insight into biological processes that are altered in CAMs and/or ATMs relative to ANMs. As a result three classes of altered processes where identified: (1) Those that are change in both CAMs and ATMs, (2) Those which show progressive changes from ATMs to CAMs and (3) those that are unique to either CAMs or ATMs. These differences are explored further in Chapter 5, however in brief: Changes seen in both CAMs and ATMs may reflect historical programming of gastric tissue myofibroblasts resulting from chronic tissue inflammation, which may have occurred prior to tumour development. In contrast, progressive changes seen between ATMs and CAMs may reflect distance related gradation in tumour mediated myofibroblast reprogramming. Finally, processes that appear to be uniquely altered in CAMs or ATMs may reflect the balance between dominance of tumour-induced changes or the suppressive effects of normal tissue stroma.

Potentially important trends were highlighted by this analysis, several of which are developed further in Chapter 5. In brief: A wide range of metabolic and inflammatory processes were altered in both CAMs and ATMs. These observations would be consistent with a reverse Warburg type system in which tumours program stromal myofibroblasts to provide nutrients for tumour growth. One potentially interesting finding was the fact that the TGF-β induced EMT pathway was uniquely over-represented in CAMs. Interestingly, expression of TGF-β2 itself was also increased in CAMs. As TGF-β has been shown to 'activate' fibroblasts (Untergasser

et al., 2005) it is possible that CAMs may play a role in inducing ANM to ATM or ATM to CAM conversion. CAMs also showed a very clear Helicobacter pylori induced signalling response that was not seen in ATMs. Also, glycogen/fatty acid synthesis appears to be inhibited in CAMs while glycolysis and fatty acid oxidation is increased (Steinberg and Kemp, 2009). This observation may provide a new way of differentiating CAMs and ATMs *in vivo.* As the N-glycan pathways are significantly over-represented in CAM vs. ANM and CAM vs. ATM, but not in ATM vs. ANM comparisons, it is possible that N-glycan biosynthesis may be an important feature of tumour 'conditioned' CAMs.

# 4 Chapter Four: Network and multivariate analysis of cancer related gene expression signatures.

## 4.1 Introduction

Canonical pathways are well-established biological processes, containing defined gene members. However, it is now clear that pathways do not operating in isolation. Also, changes in different isoforms of genes, such as adenylate cyclase, can elicit different stimuli, which are all transformed through a common point among multiple signalling pathways (Jordan et al., 2000) (Figure 4.1).

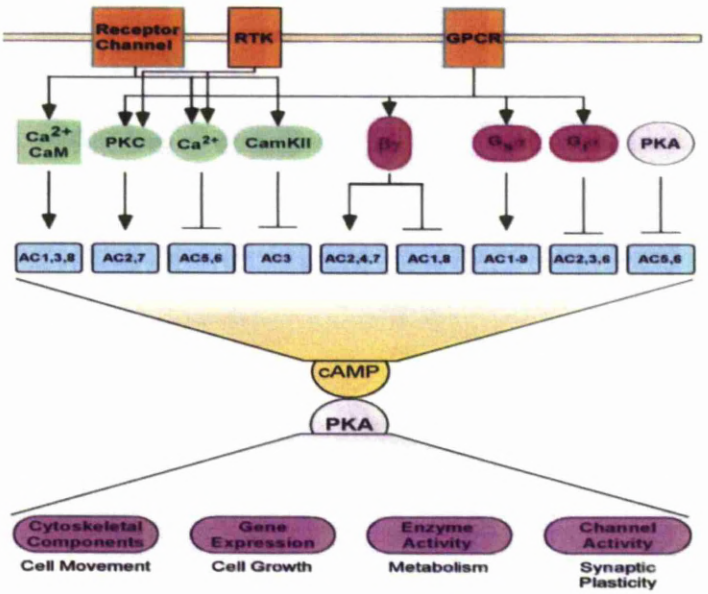



**Figure 4.1.** Many isoforms of adenylate cyclase act through PKC to produce a range of different intracellular downstream effects (Jordan et al., 2000).

Therefore, it is now clear that many genes may operate as components of multiple pathways (Figure 4.2). However, conventional pathway enrichment algorithms do not reflect the potential impact that a change in expression of a single pathway component could have on multiple pathways; or if it represents an essential part of a single pathway, or the junction or hub point between multiple pathways. The aim of the work described in this chapter is to go beyond the use of basic pathway enrichment methods, to investigate the concept of load and hub perturbation effects on processes within integrated biological networks. It is clear from analysis of any interactome network that crosstalk between pathways must be both extensive and regulated. In addition, most studies to date have not addressed the potential issue of load, or accumulative burden within particular biological processes. For example use of fold change thresholds to identify 'significant' changes may mask the accumulative effects of small apparently less-significant

changes in multiple components of the same pathway. This concept can be extended to ask what would be the added effects of multiple changes in proteins not annotated as conventional pathway members but which interact with one or more pathway components? Could changes in these components also contribute to phenotypic cause effect relationships? Here the human protein-protein interactome is used to increase coverage of differentially regulated genes, which may affect one or more known biological processes. In addition, statistical techniques have been applied to try to detect affected proteins that participate in multiple over-represented processes.

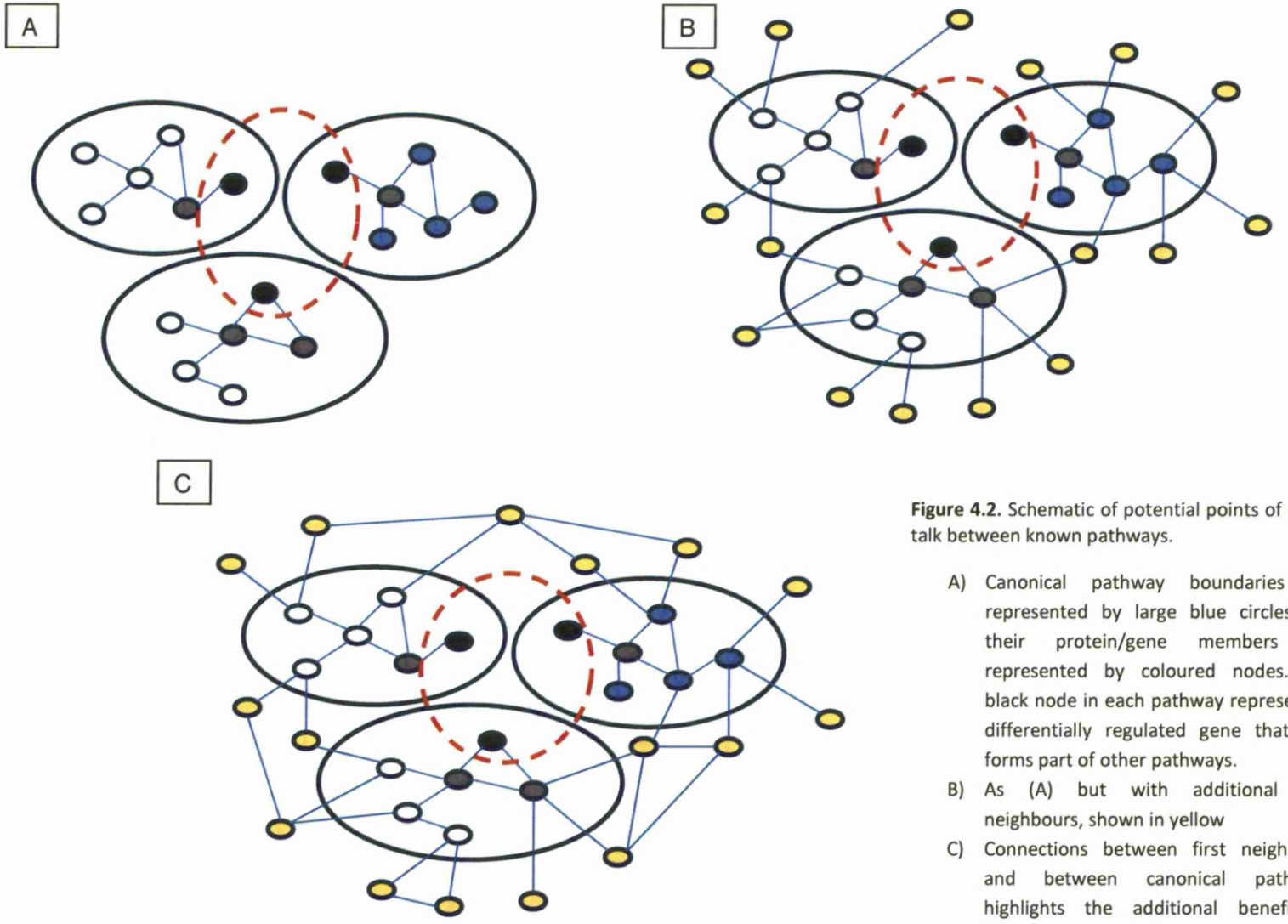



**Figure 4.2.** Schematic of potential points of cross-talk between known pathways.

A) Canonical pathway boundaries are represented by large blue circles and their protein/gene members are represented by coloured nodes. The black node in each pathway represents a differentially regulated gene that also forms part of other pathways.

B) As (A) but with additional first neighbours, shown in yellow

C) Connections between first neighbours and between canonical pathways highlights the additional benefit of interaction networks.

## 4.2 Results

### 4.2.1 Preliminary network analysis findings

Network analysis was performed on the available dataset, which was later refined in accordance with new patient information, however the underlying principles, and observed trends provide evidence to support the idea that this form of network analysis may provide additional insight into the potential significance of the large number of significantly changed genes that are not assigned to pathways by conventional pathway enrichment methods.

Genes mapped to Metacore™ pathways were mapped onto the human interactome, this revealed a high level of connectivity between proteins relating to genes with altered gene expression profiles (Figure 4.3A). Although the genes fall into clear pathways when utilizing pathway enrichment tools such as Metacore, there is no distinction of pathways when significantly changed genes are mapped into the human interactome. This trend was investigated using an in-house network analysis tool called 'Hypernode' (Figure 4.3B). In essence, the Hypernode algorithm condenses all genes within a certain pathway onto a single node. Therefore, connections between nodes represent connections between genes present within different pathways. This demonstrates that although we tend to think of pathways as having linear order, they are obviously not working in isolation. They are highly connected processes often sharing multiple components. As such, disruption of shared components can potentially have more dramatic effects on multiple pathways and the system as a whole.

In addition, it is clear that genes that were not mapped to Metacore pathways, form highly connected networks (Figure 4.3C). Such genes would have been ignored using standard canonical pathway tools, but are clearly highly connected. Therefore, it is important to investigate which genes they may be interacting with and what pathways their differential expression could potentially effect.

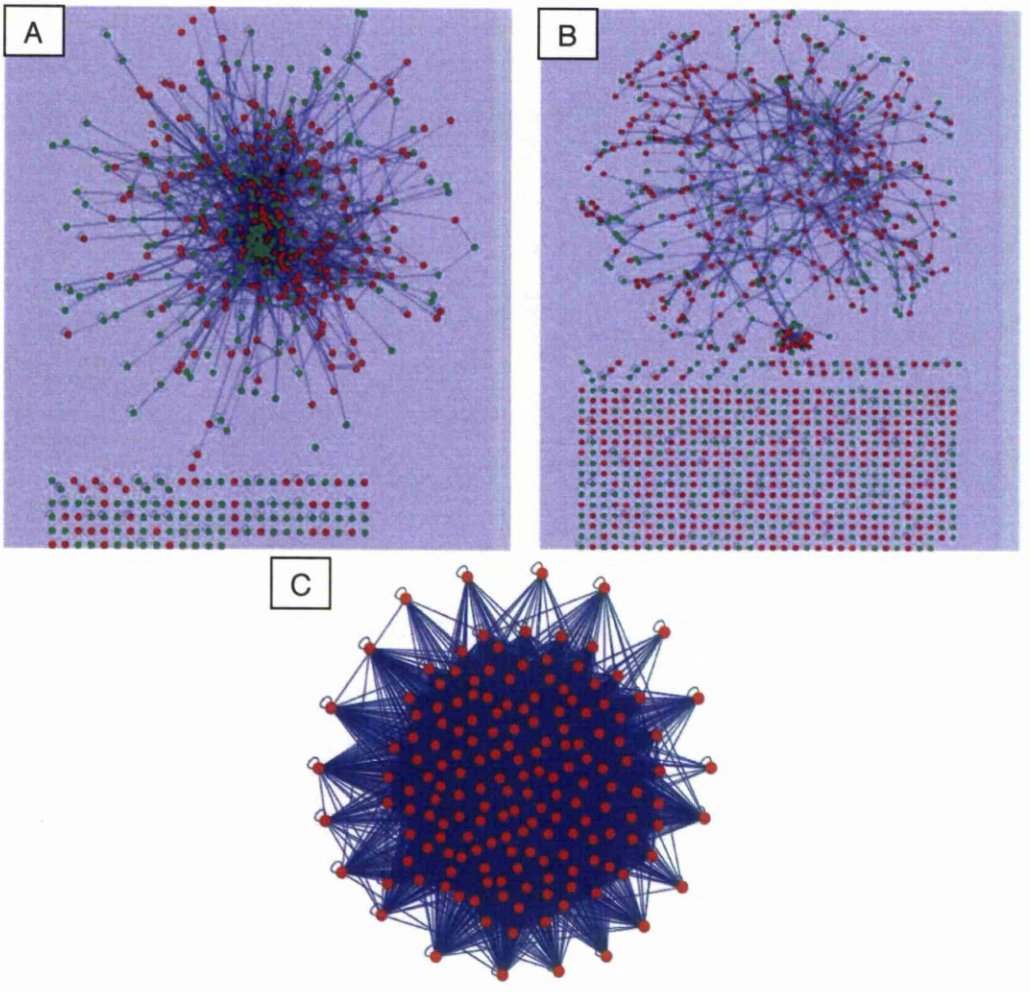



**Figure 4.3.** Protein-protein interaction networks A) Nodes represent differentially regulated genes mapped to Metacore pathways and connections represent the interactions between them B) Nodes represent differentially regulated genes unable to be mapped to Metacore pathways and connections represent the interactions between them. Within A) and B) red nodes represent up regulated genes and green nodes represent down regulated genes. C) Hypernode network, nodes represent pathways and the connections between pathways represent interactions between genes, red coloured nodes do not represent direction of fold change.

Visualisation of pathway cross-talk/connectivity via the human interactome demonstrates the complex nature of connectivity and potential cross-talk between pathways, and the need for new techniques that can highlight multi-functional/connecter genes. In theory, the identification of genes that play a role in many pathways, is important, as the differential regulation of a small number of promiscuous genes could alter cause/effect relationships across a wide range of canonical pathways.

### 4.2.2 Assembling a Myofibroblast specific protein interaction network

As all genes are not be expressed in all cell types, it was important to first generate a myofibroblast specific gene expression network, in which all genes that were detected at a baseline level in microarray studies are included (Figure 4.4A). The resulting network consists of 10,959 nodes and 80,095 edges. As explained in methods section 2.6.1, this network also contains interologs, which are binary protein interactions that are predicted to be conserved between homologous proteins in another species.

From this myofibroblast network, individual sub-networks were then created for each dataset. The CAM vs. ANM network consists of 10,667 proteins, connected by 78,615 interactions, the ATM vs. ANM normal network consists of 10,768 proteins connected by 78,974 interactions and the CAM vs. ATM normal network consists of 10,722 proteins connected by 77,584 interactions (Figure 4.4A-C). For each dataset, there are clearly a large number of significantly changed genes, which are highly connected in the network, representing a high degree of communication.

Therefore, as the level of cross-talk within individual datasets is high, we aim to identify the differentially regulated processes/pathways and identify the sites of cross-talk, within these highly connected networks.

In addition, we wished to gain insight into the pathways that may be affected by the large numbers of differentially regulated genes, which themselves are not known to be core members of canonical pathways. Therefore, differentially regulated genes mapped to Metacore™ pathways were mapped into the interactome, and 1-step networks were produced, which contains all proteins that interact with differentially regulated genes. Differentially regulated genes within the one step shell were then identified. Next, we assessed whether differentially regulated genes from the one-step shell would have been assigned to canonical pathways using other pathways analysis tools, such as DAVID, Ingenuity and Reactome. For the CAM vs. ANM dataset, an additional 230 differentially regulated genes which directly interacted with differentially regulated genes mapped to Metacore pathways. Similar results were found for the other two datasets; with 292 and 174 genes in the CAM vs. ATM and ATM vs. ANM datasets respectively. Analysis of significantly changed genes within the context of binary protein-protein interaction network, have highlighted the large amount gene expression information that is lost when using canonical pathway analysis tools in isolation.

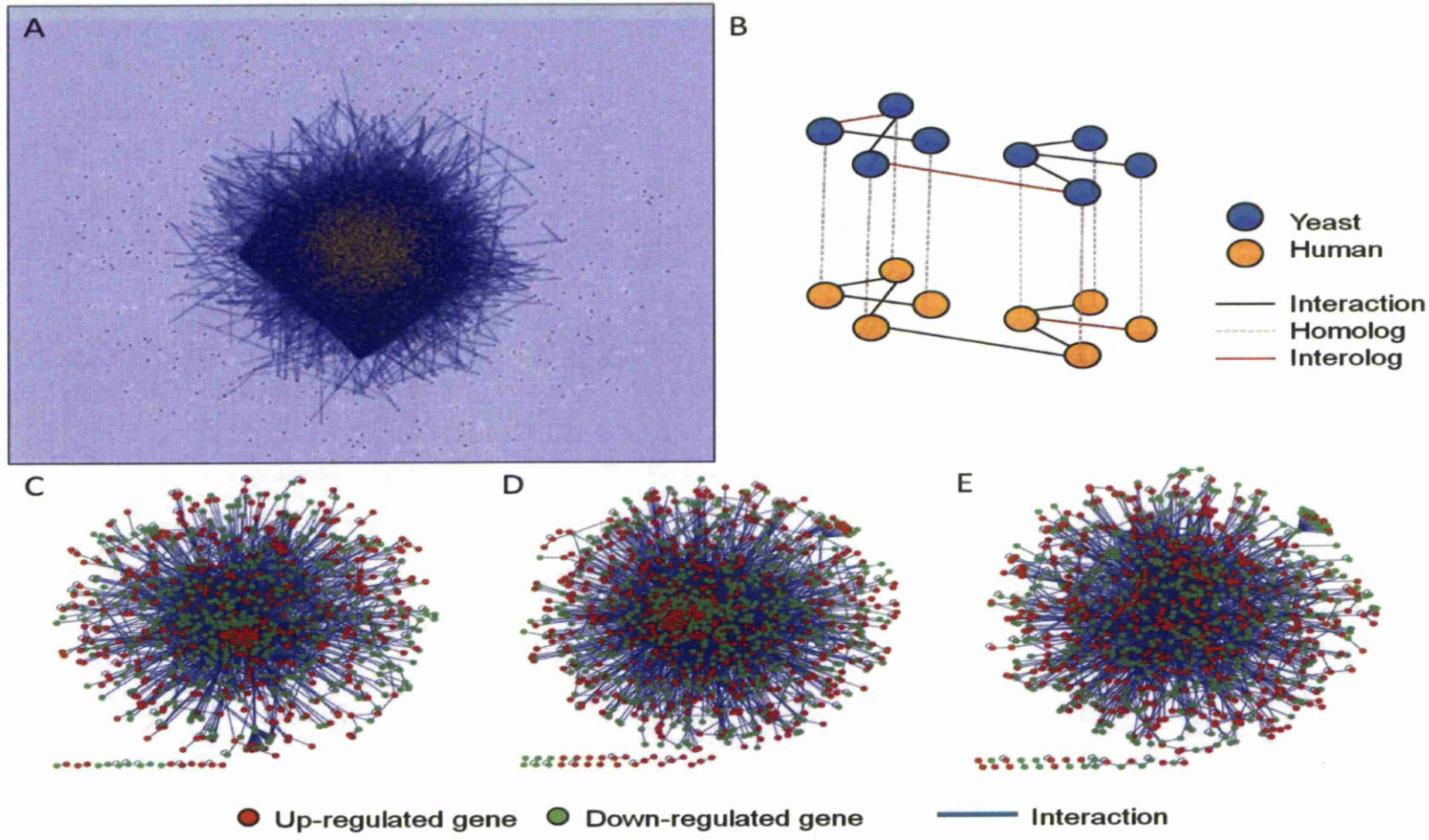



**Figure 4.4** Human protein-protein interaction networks with interologs A) Myofibroblast interaction network B) Examples of interolog interactions between homologous proteins C) CAM vs. ANM network D) ATM vs. ANM network E) CAM vs. ATM network. Proteins are represented by nodes and interactions between proteins are represented edges. Within C-E, red nodes represent up-regulated genes and green nodes represent down-regulated genes.

Extending this principle we wished to identify all differentially regulated genes that interact not just with differentially regulated genes assigned to canonical pathways but to any member of canonical pathways that contain differentially regulated proteins. This type of information is extremely hard to access from Metacore, as the software does not allow large-scale export of individually annotated GeneGo pathway information. Individual pathways can however be selected and all genes within that pathway exported, one pathway at a time. Therefore exporting all members of all pathways would be extremely time-consuming, as our average dataset hits around 500 different pathways.

### 4.2.3 Multi-variant analyses

Unlike the hypothesis driven tests explained above multi-variant analysis techniques are exploratory, providing data from which hypothesises can then be derived. Multi-dimensional scaling is used to form groups within the data, where a group represents a sub-section of data that is more similar to each other.

#### 4.2.3.1 Multi-dimensional scaling

For this sub-section, all of the analysis presented was performed for each dataset, but as the results from each were so similar, only example figures from the CAM vs. ATM dataset are shown as an example. Figure 4.5, shows different techniques commonly applied to detect over-represented pathways or processes within a given gene list: the Fisher exact test, the 2x2 Chi-Squared test and combined 3x2 and the Chi-Squared test. Each Reactome pathway is plotted against its over-representation p-value for each given test and a dotted line is placed at a p-value of 0.05, therefore

any pathways plotted below this threshold are classed as statistically over-represented. It is apparent that each statistical test gives a completely different pattern of pathway over-representation, with a very different number of pathways being classed as over-represented. The application of the 3x2 Chi-squared test rather than the 2x2 Chi-squared test also dramatically alters the interpretation of the results. Below is an example of the 3x2 contingency table formed to carry out such statistical tests:

| | Pathway | |
|---|---|---|
| | Yes | No |
| Diff. Expressed | a | d |
| Background | b | e |
| Not in background | c | f |

The decision of which test to use is complicated, a 2x2 contingency table is formed using only a, b, d and e, not taking into account genes not in the background set. The inclusion or exclusion of genes not in the background dramatically affects the results (Figure 4.5). In this analysis we are looking for differences between differentially and non-differentially expressed genes, not those genes 'not in background'. Therefore, this extra category could mean that you get a very significant P-value, but not for the difference you are looking for. The Fisher test is exact, and is used when the data distribution is known, it is very powerful when the distribution is assumed correctly but is harder to compute than the Chi-sq. The chi-squared test makes a distribution assumption; therefore larger sample sizes are

needed to ensure the distribution assumption by Chi-squared test is correct. Therefore, the decision of whether to use a Fisher or Chi-squared test also depends on the observed numbers in the contingency table, but these numbers differ for each pathway.

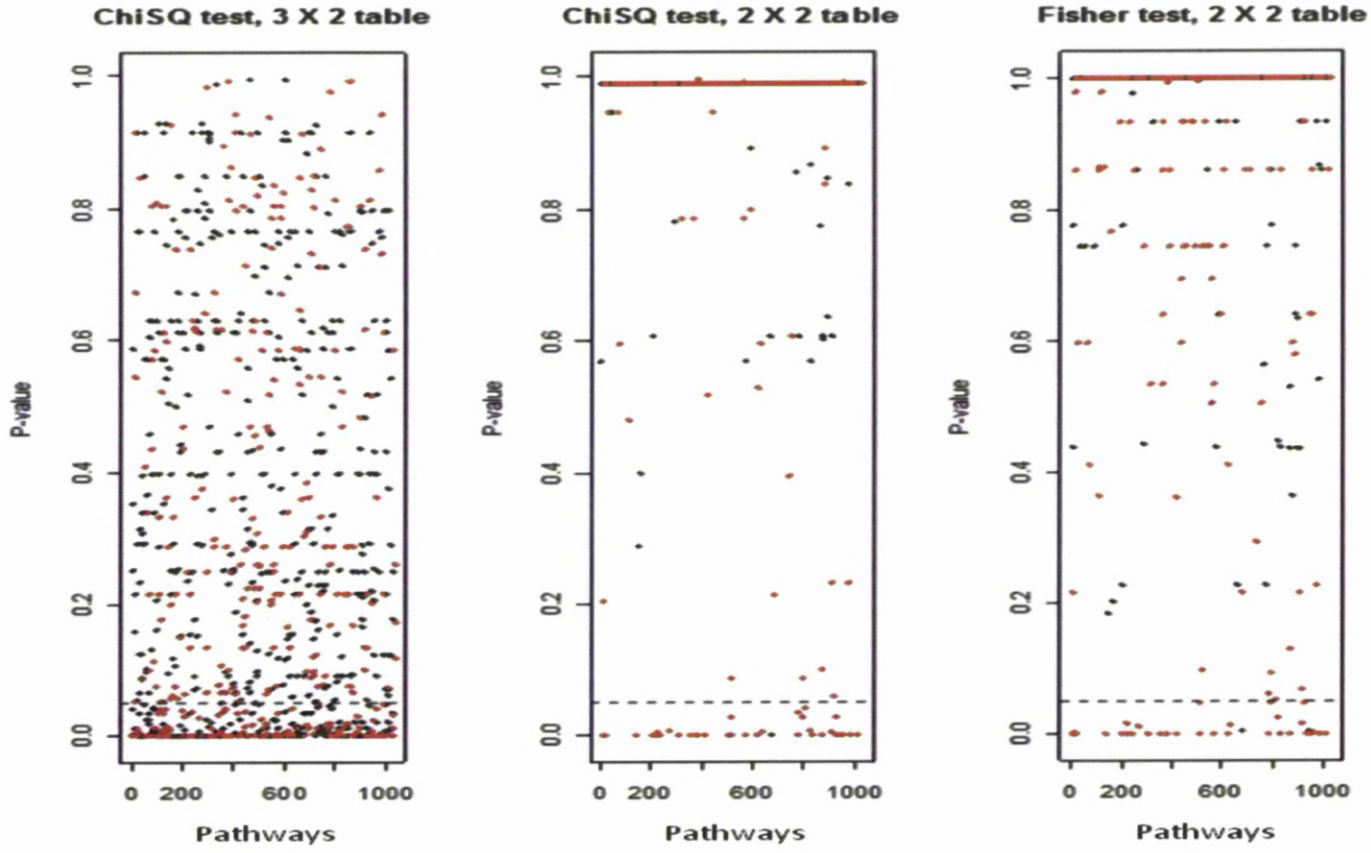



**Figure 4.5.** Three types of statistical tests frequently applied to define pathway over-representation. Black points represent pathways that have odds ratios <2 and red points represents pathways with odds ratios ≥2. The black dotted line represents an over-representation pathway of ≤0.05.

The odds ratio is suggested as a more appropriate scoring system to apply to define which pathways are over-represented, as the odds ratio does not have to make any assumptions as we are not testing anything – it is simply a statistic (like the mean, median or variance).

$$\text{odds ratio} = \frac{a*e}{b*d}$$

The odds ratio takes into account the proportion of differentially expressed genes in a pathway to the number of non-differentially expressed (background) genes in a pathway. Therefore, if an odds ratio is greater than 1 the proportion of differentially expressed genes in a pathway is greater than the proportion of background genes in a pathway. The components of the odds ratio are plotted in Figure 4.6A, when the numerator is greater than the denominator the pathway is classed as over-represented and coloured red.

Looking back to the Chi Squared and Fisher tests in Figure 4.5, black pathways are pathways with odds ratios less than 2 and red pathways have odds ratios greater than 2. It is clear that there are differences between pathways classed as over-represented, as defined using the different approaches. To try to explain this finding, the odds ratio equation needs to be broken-down; as '**e**' is expected to be at least 3 times the size of '**d**', for an odds ratio of 1 or greater, '**a**' would have to be greater or equal to b/3. Red dots with insignificant p-values are due to pathways with small numbers of background genes, therefore when '**b**' is small (i.e. <3), '**a**' only has to be equal to 1 to have an odds ratio>1. Odds ratios are calculated for each pathway in the Reactome database (Figure 4.6B), to exclude noise all odds ratios ≥2 are classed as over-represented

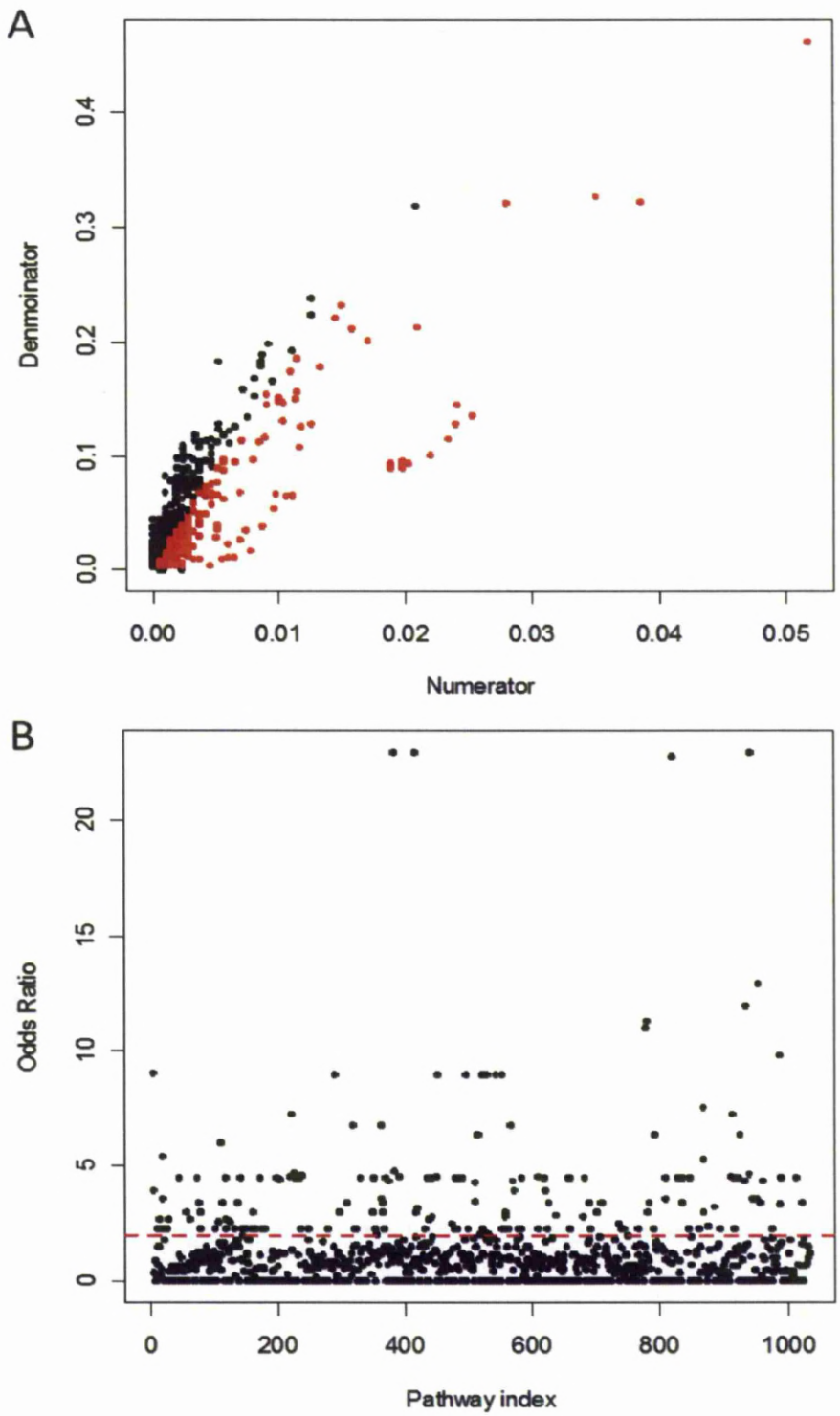



**Figure 4.6.** A) Odds ratio components, red pathways represent Reactome pathways with Odds ratios >1. B) Odds ratios for each pathway in the Reactome database. Dashed line represents an odds ratio threshold of 2.

#### 4.2.3.1.1 Similarities based on pathways

As discussed in the introduction the thought of pathways acting in isolation is physiologically unfeasible. Therefore the aim of this study was to identify groups of pathways, which are similar to one another, as they contain similar gene members. Identification of such sub-groups of related pathways is the first step in revealing groups of over-represented biological pathways, which share common components and represent areas of potential cross talk within the larger biological system. Initial hierarchal clustering analysis, of the pathway similarity analysis revealed an extremely complex dendrogram (Figure 4.7). As identification of pathway groups is impossible in this format, data was sub grouped using multi-dimensional scaling.

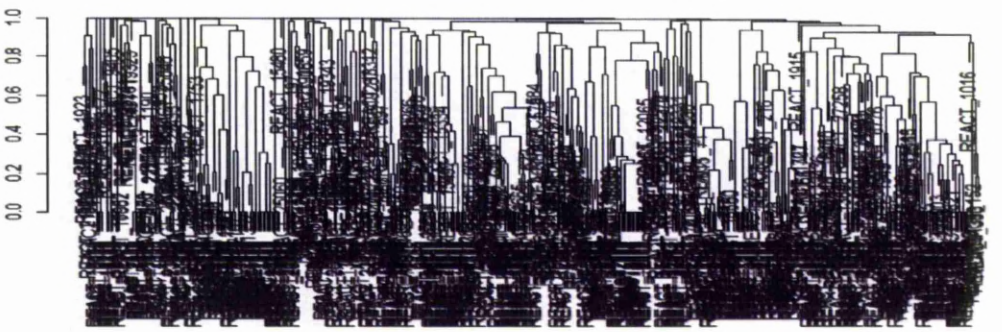



**Figure 4.7.** Resultant dendrogram from initial Hierarchal clustering analysis, similarities based on pathways. Presented to demonstrate the difficulty in identifying similar pathways without the identifaction of smaller subsets. Application of such an algorithm using small clusters identified using Multi-Dimensional Scaling would be informative.

As multi-dimensional scaling uses the matrix of dissimilarities, the pathways are plotted based on their dissimilarities to one another and the dissimilarities are represented by geometric distance. Multidimensional distances are plotted in four dimensions, the CAM vs. ANM dataset is shown as an example (Figure 4.9). Identification of pathways more similar to one another, in a 4-dimensional space, can be confusing. Therefore, pathways were generated as interactive 3-dimensional

plots to aid subsequent analysis/visualization. For clarity, all datasets were visualised as 3D plots in which only over-represented pathways are visualised, a screen shot of the CAM vs. ANM normal 3D multi-dimensional plot is shown in Figure 4.8. Using these interactive 3D clusters plots, highly interconnected pathways were identified and highlighted for investigation in the correct 2-dimensional plane (Figure 4.10)

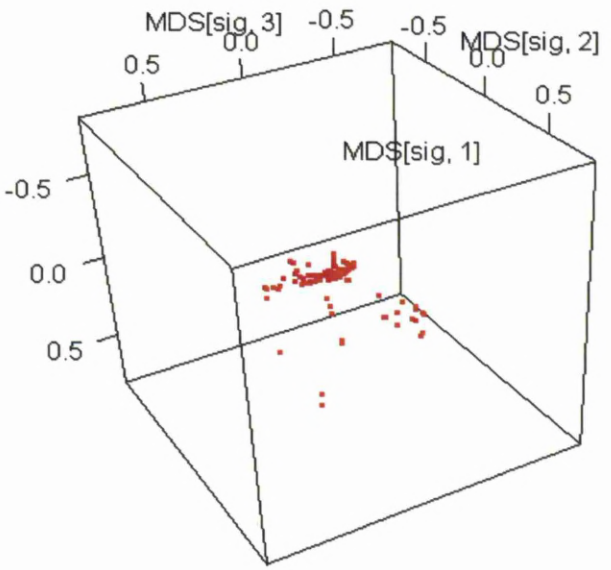



**Figure 4.8**. CAM vs. ANM dataset. Screen shot of an interactive 3D MDS plot, based on the similarity of pathways. Only over-represented pathways are displayed as red dots.

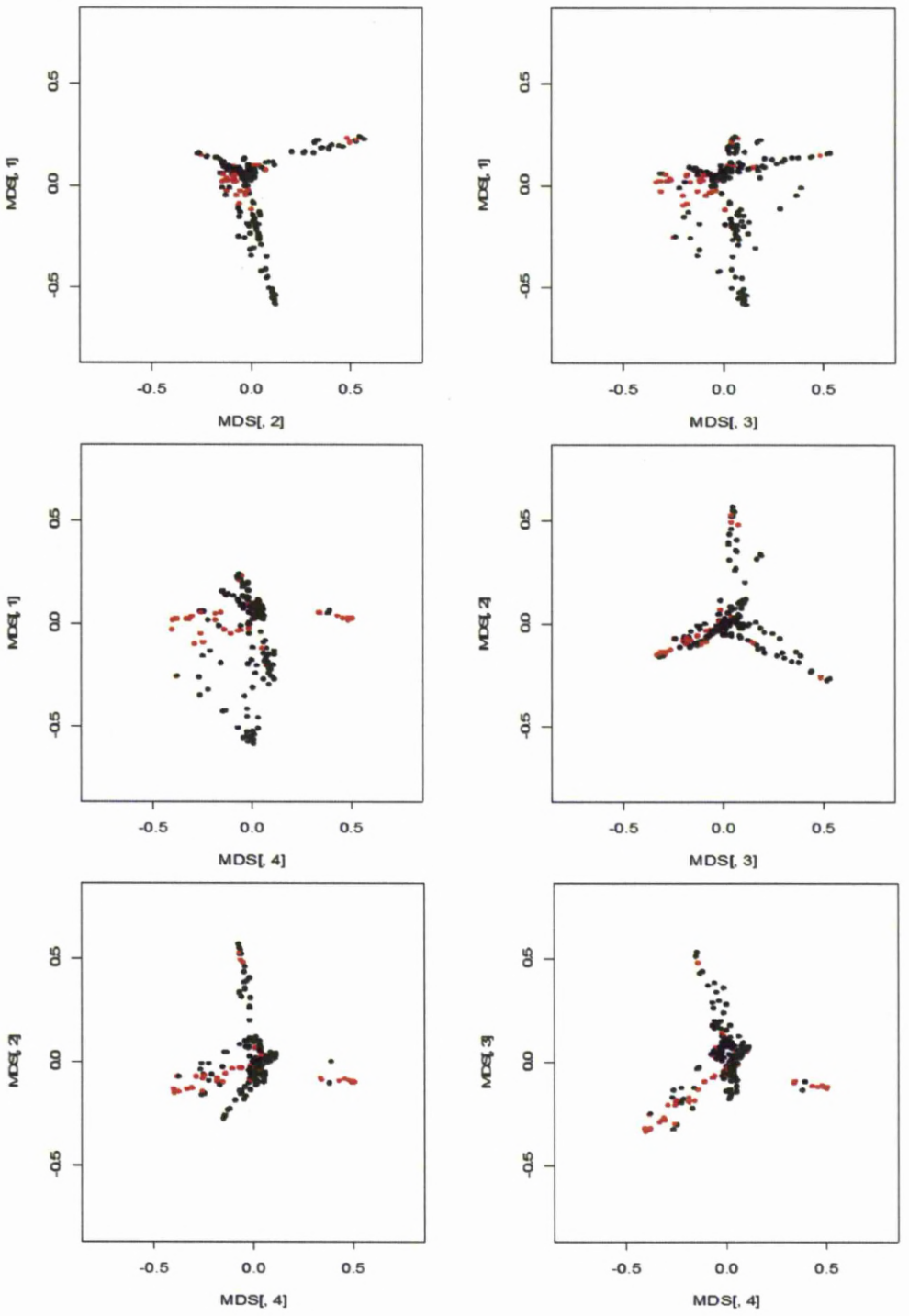



**Figure 4.9.** Multi-dimensional scaling based on the similarity of pathways, visualised in four dimensions. Non-overrepresented pathways are represented by black points and overrepresented pathways (Odds Ratio>2) are represented by red points. MDS number as labelled on axis represents the dimensions visualised.

Although this study is specifically focused on over-represented pathways, all pathways need to be taken into account when calculating the multi-dimensional co-ordinates, as excluding them could affect how dissimilar the over-represented pathways appear to one another. After calculation of co-ordinates with all pathways included, only over-represented pathways are visualised to helped distinguish clusters of over-represented pathways observed from the CAM vs. ANM (Figure 4.10A&B), CAM vs. ATM (Figure 4.10C&D) and ATM vs. ANM (Figure 4.10E&F) datasets.

#### 4.2.3.1.1.1 CAM vs. ANM

Six clusters of pathways were identified in the CAM vs. ANM dataset, all in the 1st and 3rd dimensions (Figure 4.10B). Pathways within each cluster are listed in Supplementary File 4.1. Cluster 5 is the large central cluster of pathways and therefore represents potentially the largest site of cross talk between over-represented pathways, which was the aim of applying Multidimensional scaling and is therefore analysed in detail below. The other clusters represent areas of less dense clusters and therefore less potential cross-talk, therefore although these pathways are over-represented they provide less indication of the small numbers of differential regulation effecting large numbers of pathways/processes, a detailed analysis of such clusters have been analysed and provided as a resource in supplementary chapter 1, section 7.1.1.1.1.

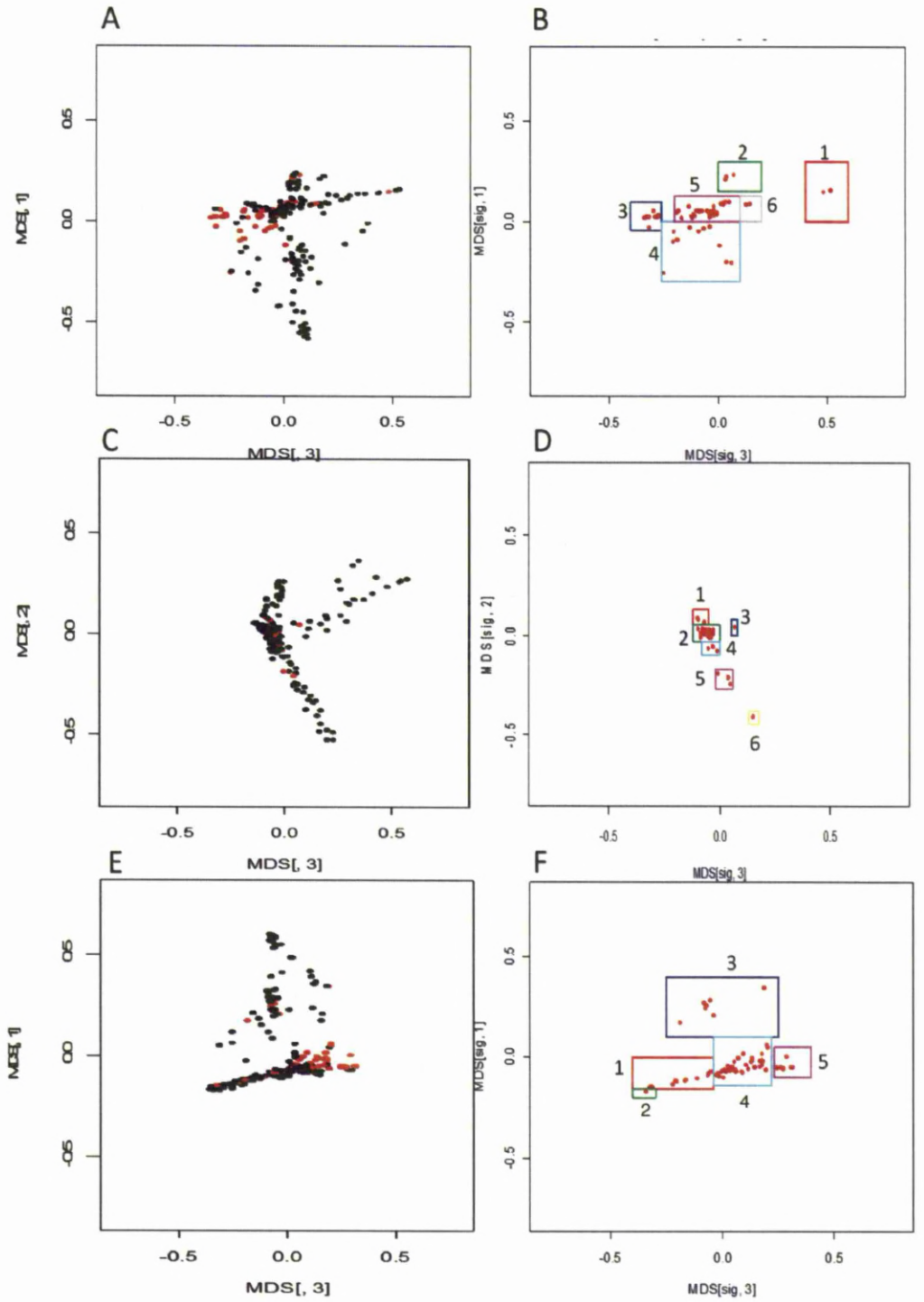



**Figure 4.10**. Multi-dimensional plots based on the similarity of pathways. CAM vs. ANM (A and B). CAM vs. ATM (C & D). ATM vs. ANM (E & F). Black dots indicate non-over-represented pathways and red dots indicate over-represented pathways with odds ratios >2. Clusters of pathways are identifiable by coloured boxes. MDS number as labelled on axis represents the dimensions visualised.

Dense cluster 5 represents a range of diverse processes. A subset of over-represented pathways include immune related processes, with a large proportion of differentially regulated genes being up-regulated in interferon signalling and interactions between lymphoid and non-lymphoid cells. Also, within this central cluster we found a clear group of over-represented pathways involved in energy metabolism, specifically beta-oxidation. All differentially regulated genes involved in these pathways were up regulated enzymes, suggesting an increase in energy generation, through fatty acid oxidation via the mitochondrial beta-oxidation pathway.

Interestingly, multiple pathways involved in the transport of energy metabolites across the cell membrane were also found to be differentially regulated. These include two over expressed facilitated glucose transporters; solute carrier family 2 member 10 and solute carrier family 2 member 11. In cancer, high levels of cell proliferation requires large amounts of energy, therefore glucose is actively imported into the cell and glucose transporters are often up-regulated (Macheda et al., 2005). The differentially regulated proton-coupled monocarboxylate transporter, solute carrier family 16 member 3 was also found to be over-expressed. This class of transporters are essential for the transport of monocarboxylates, such as pyruvate and lactate across the cell membrane (Morris and Felmlee, 2008). Finally, zinc and bile salt transporters were also up-regulated. Remaining pathways include those affecting cell life or death, such as the cell survival promoting nuclear kinase AKT pathway are negativity regulated, with AKT itself and CREB both being down-regulated. Pathways involving the regulation of apoptosis are over-represented, two pro-apoptotic BH3 proteins BID and BAD are

up regulated, whilst the anti-apoptotic protein BCL-2 is also increased (twice as much as BID and BAD) suggesting that on balance cell survival most probably remains stable, which agrees with levels of cell survival when cultured *in vitro*.

#### 4.2.3.1.1.2 CAM vs. ATM

Six clusters of pathways were identified in the CAM vs. ATM dataset, all in the 2nd and 3rd dimensions (Figure 4.10D). Pathways within each cluster are shown in Supplementary File 4.2. As for the other datasets the dense cluster (cluster 2) is analysed in detail here and the other clusters are analysed and provided as a resource in Supplementary chapter 1, section 7.1.1.1.2.

Cluster 2 represents by far the largest cluster of connected pathways, including the unfolded protein response pathway, which contains the largest amount of differentially regulated genes, the pathway is a common cellular stress response pathway, triggered by the mis-folding or unfolding of proteins within the endoplasm reticulum. Pathways involved in interferon signalling are also over-represented. Interferons are released by host cells upon the detection of tumour cells, activating an immune response. Interferon's along with several antigen presenting proteins (HLA's) are up-regulated within the myofibroblasts, suggesting increased inflammation and a heightened immune response. Genes involved in mTOR signalling are down regulated. mTOR up-regulation has been associated with cell proliferation, angiogenesis, inhibition of apoptosis and interestingly aerobic glycolysis (Sun et al., 2011). Several other over-represented pathways are involved in LKB1 regulating AMPK and AMPK inhibiting mTOR, all differentially regulated genes in these over-represented pathways are down regulated. Generally, over-

represented apoptotic pathways have many down regulated genes, including caspases suggested decreased programmed cell death.

Similar to the CAM vs. ATM dataset, pathways involved in fatty acid oxidation are also over-represented. The pathway; Activated AMPK stimulates fatty-acid oxidation in muscle is over-represented, activation of AMPK stimulates fatty acid oxidation and inhibits fatty acid synthesis, genes involved in this pathway are both up and down regulated. Enzymes involved in cholesterol biosynthesis are down-regulated, suggesting the end product of fatty acid β-oxidation, acetyl co-A, enters the mitochondria to feed into the citric acid cycle, rather than being used to drive the biosynthesis of cholesterol. In addition, as seen in the CAM vs. ATM dataset, a number of solute carriers pathways are over-represented, the mono-carboxylate transporter SLC16A7 is under expressed, whilst SLC16A3 is over-expressed.

Genes involved in basigin (BSG) interactions are down-regulated. BSG is a matrix metalloproteinase inducer, which themselves degrade the extracellular matrix. The differentially expressed genes in this pathway have mixed expression, several solute carriers are over-expressed whist the integrin B1 is under-expressed. In many cancers, integrin expression is increased in cells with metastatic potential (Cooper et al., 2002). Finally, differentially regulated genes involved in other cell-extracellular matrix interactions over-represented pathways involved in cell-extracellular matrix interactions are also down regulated.

#### 4.2.3.1.1.3 ATM vs. ANM

Five clusters of pathways were identified in the ATM vs. ANM dataset, all in the 1st and 3rd dimensions (Figure 4.10F). Pathways within each cluster are shown in Supplementary File 4.3. As for the other datasets the dense cluster (cluster 4) is analysed in detail here and the other clusters are analysed and provided as a resource in Supplementary chapter 1, section 7.1.1.1.3.

Within the dense cluster 4, there seems to be a selection of pathways that fall within metabolism of proteins, specifically translation, initiation, processing and termination. The majority of such differentially regulated genes are up regulated, indicating an increase in protein production. A range of pathways involving cell cycle checkpoints, including the activation of ATR in response to replication stress pathway. Normal activation of this pathway would inhibit DNA replication and initiate DNA repair, therefore the down regulation observed suggests initiation of DNA replication and inhibition of DNA repair, supporting other translation pathways which have increased expression of differentially regulated genes. All genes involved in E2F1 regulated gene expression, the inhibition of replication pathway and DNA glycosylases (enzymes involved in base excision repair) are decreased. E2F1 usually inhibits gene expression (Iglesias-Ara et al., 2010) and DNA glycosylases are enzymes, which are involved in DNA base excision repair, down regulation of pathways such as these suggest a lack of checkpoint control and of cellular proliferation.

All genes involved in metabolism of amino acids, carbohydrates and fatty acids are increased. Specifically for the metabolism of carbohydrates, the pentose phosphate pathway is over-represented and for the metabolism of fatty acids, the β-oxidation

pathway is over-represented. Remaining pathways include decreased cell junction organisation and cell-extracellular interactions and pathways involving the regulation of apoptosis are over-represented, both pro-apoptotic BH3 proteins BID and BAD are up regulated, suggesting an increase in apoptotic cell death but as mentioned previously these cells are not dying and therefore most probably opposing pathways are over-represented within alternative clusters.

#### 4.2.3.1.1.4 Comparison of Dense Pathway Clusters Across Datasets.

A similar range of pathways are over-represented and present within the central dense clusters across datasets. Common pathways include metabolic pathways, those involved in fatty acid β oxidation, cholesterol synthesis and transport of metabolites across the cell membrane. All datasets show an increase in fatty –acid oxidation but cholesterol biosynthesis is only decreased in the CAM vs. ATM dataset, suggesting that it is specifically lost CAMs. Interferon signalling is increased in CAM vs. ANM and the CAM vs. ATM datasets, suggesting an increase in inflammation in the microenvironment of cancer but not ATM myofibroblasts.

### 4.2.3.1.2 Similarities based on genes

The aim of this part of the analysis was to identify groups of genes that are similar to one and other, due to the range of pathways in which they are members. Identifying genes that are present in many over-represented pathways is important, as they are the genes that represent sites of cross-talk between pathways. Therefore, if differentially expressed they have the potential to affect large numbers of processes. Identification of such promiscuous genes may therefore be important,

as they may be responsible for producing the range of biological processes that result in the cancer myofibroblast phenotype.

As for the similarity based on pathways, initial hierarchal clustering analysis, of gene similarity revealed an extremely un-clear dendrogram, as the dataset is far too large (data not shown). Identification of groups of similar genes was impossible and the data needed to be sub grouped using multi-dimensional scaling. Genes are plotted based on their dissimilarities to one another, the dissimilarities represented by geometric distance, genes are first visualised as interactive 3D plots enabling clusters to be identified within the correct 2-dimensions. For all three datasets, six clusters of genes were identified in the first three dimensions (Figure 4.11). When scrolling through the lists of genes differentially regulated in myofibroblasts derived from different regions, as the lists are so extensive a starting point is to look at genes which take part in a larger number of differentially regulated pathways, as they have the potential to cause most disruption if altered.

#### 4.2.3.1.2.1 CAM vs. ANM

Six clusters of genes were identified in the CAM vs. ANM dataset, two in the 1$^{st}$ and 2$^{nd}$ dimensions and four in the 1$^{st}$ and 3$^{rd}$ dimensions (Figure 4.11 A and B). Genes within each cluster are shown in supplementary excel file 4.4. Cluster 6 is the largest central cluster of genes, with genes present in the largest number of over-represented pathways and therefore represents potentially the largest site of cross talk between pathways, which was the aim of applying multidimensional scaling and is therefore analysed in detail below. The other clusters represent areas of less dense clusters and therefore less potential cross-talk, therefore although these genes are differentially regulated they provide less indication of the small numbers

of differential regulated genes effecting large numbers of pathways/processes, a detailed analysis of such clusters have been analysed and provided as a resource in Supplementary Chapter 1, section 7.1.1.2.1. Cluster 6 represents the largest cluster of 296 significantly changed genes, 131 of which provide cross talk between 124 over-represented pathways (OR<2). The majority of genes in multiple pathways are those involved in the DNA polymerase factors, elongation factors, DNA repair, DNA replication, cell cycle genes, pro-apoptotic genes, transcription factors and those involved in the MAPK/ERK pathway. The proliferating cell nuclear antigen (PCNA) is down-regulated and play a role in 25 different pathways, it is a clamp protein helping to hold DNA polymerase to the DNA.

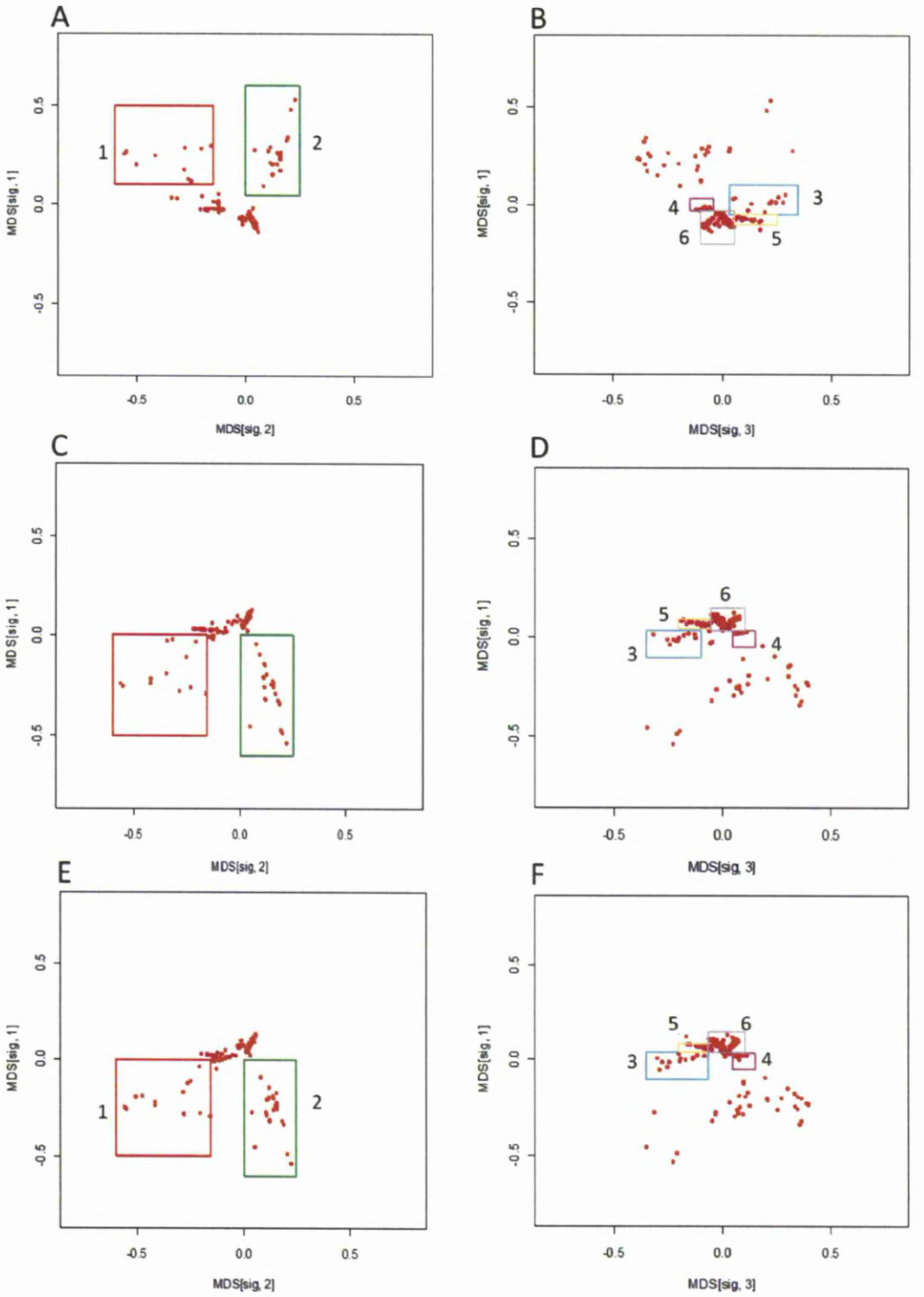



**Figure 4.11**. Multi-dimensional plots based on the similarity of genes A and B) CAM vs. ANM. C and D) CAM vs. ATM. E and F) ATM vs. ANM. Black dots indicate non-over-represented pathways and red dots indicate over-represented pathways with odds ratios >2. Clusters of pathways are identifiable by coloured boxes. MDS number as labelled on axis represents the dimensions visualised.

#### 4.2.3.1.2.2 CAM vs. ATM

Six clusters of genes were identified in the CAM vs. ATM dataset, two in the 1st and 2nd dimensions and four in the 1st and 3rd dimensions (Figure4.11 C and D). Genes within each cluster are shown in Supplementary excel file 4.5. As for the other datasets, the dense cluster (cluster 6) is analysed in detail here and the other clusters are analysed and provided as a resource in Supplementary chapter 1, section 7.1.1.2.2.

Cluster 6 is by far the largest cluster, a group of 303 significantly changed genes, 149 of which provide cross-talk between 110 over-represented pathways (OR<2). Top most promiscuous genes include, several types of integrin, both of which are down-regulated, a multitude of signalling cascade intermediates, and the fibroblast growth factor rector which is also down-regulated. Therefore, the majority of over-represented pathways include signalling pathways, cell-cell and cell-extracellular interaction pathways, transmembrane transport of a wide selection of biological molecules and apoptosis

#### 4.2.3.1.2.3 ATM vs. ANM

Six clusters of genes were identified in the ATM vs. ANM dataset, two in the 1st and 2nd dimensions and four in the 1st and 3rd dimensions (Figure4.11 E and F). Genes within each cluster are shown in Supplementary excel file 4.6. As for the other datasets, the dense cluster (cluster 6) is analysed in detail here and the other clusters are analysed and provided as a resource in Supplementary chapter 1, section 7.1.1.2.3.

Cluster 6 again represents by far the largest group of genes, with 374 significantly changed, of which 179 provide cross-talk between 131 over-represented pathways (OR<2). The top most promiscuous genes are involved in DNA translation/regulation/repair, chromosome maintenance and multiple signalling pathways. In addition, there is a set of over-represented pathways involved in fatty acid metabolism, with all genes are up-regulated, and although they are by far not the most connected interconnected pathways, mitochondrial β-oxidation enzymes show crosstalk between a maximum of 7 over-represented metabolic pathways.

### 4.2.4 Network Modularity – Netbox

The online resource Netbox (Cerami et al., 2010), was used to identify similar modules within the CAM and ATM myofibroblasts, at different stages of the disease, to distinguish driver from passenger genes. Differentially regulated genes were connected either directly or through a single linker node (an un-altered gene that directly connects two altered genes) to produce a network. Linker genes with $p \leq 0.05$, are more highly connected than would be expected by random chance within the same network, and were therefore included in the network.

The CAM vs. ANM dataset generated a network of 711 differentially regulated genes and 325 linker genes, resulting in 142 modules (47 modules with ≥4 gene members). Details of all modules and linker genes are given in Supplementary File 4.7. An observed local network modularity score of 0.615088 and a random local modularity score of 0.006481 resulted in a scaled network modularity score of 307.876. Therefore, showing the CAM vs. ANM network demonstrates more local modularity than would be expected by random chance. Previous analysis in chapter

3 shows that 818 differentially regulated CAM vs. ANM genes were unable to be mapped to pathways by any canonical pathway tool. Using this algorithm, 156/818 are mapped onto modules using the Netbox software tool. A selection of the largest detected modules are shown in Figure 4.12, all of which contain genes which were unable to be mapped to canonical pathways. Biological processes relating to each module were retrieved using Cytoscapes BinGo plug-in. The biological processes of the CAM vs. ANM modules include; Response to stimulus/DNA damage; Regulation of metabolic processes; Nucleotide metabolic processes; Fatty- acid oxidation/energy metabolism; Apoptosis/cell death; DNA replication/DNA repair; Transcriptional regulation and Cell cycle/cell division, mitosis and M-phase. The two largest modules, relating to RNA processing and translation, are not shown, as their large size would affect the clarity of the figure and they did not incorporate any genes that were un-mapped to canonical pathways.

The ATM vs. ANM dataset generated a network of 835 differentially regulated genes and 189 linker genes, resulting in 245 modules 77 modules ≥4 gene members). Details of all modules and linker genes are given in Supplementary File 4.8. An observed local network modularity score of 0.611947 and a random local modularity score of 0.004807 resulted in a scaled network modularity score of 290.317. Therefore, like the CAM vs. ANM dataset, the ATM vs. ANM network also demonstrates more local modularity than would be expected by random chance. 200/1049 differentially regulated genes, which were unable to be mapped to canonical pathways, are included in modules using this Netbox software tool. The 6 largest modules containing genes that were un-mapped to canonical pathway are

shown in Figure 4.13. The ATM vs. ANM modules biological processes include; Signal transduction, Response to stimulus/DNA damage; DNA replication/DNA repair; Transcriptional regulation; Cell cycle/cell division/mitosis and M-phase cellular organelle organisation. Importantly, clusters of modules in the CAM vs. ANM and ATM vs. ANM datasets reveal real biological processes, demonstrating that modules detected within the network are not just an artefact of biological networks, as suggested previously (Hallinan, 2004). Comparison of modules detected in both datasets also reveals modules involved in similar biological processes. The CAM vs. ATM dataset generated a network of 648 differentially regulated genes and 102 linker genes. However, the modules were surprisingly weak, only 4 modules came through, with the largest containing over 400 genes. Details of all modules and linker genes are given in Supplementary File 4.9. Upon analysis, this large module contained a range of very different biological pathways, from transcription, chromosomal maintenance to fatty acid metabolism. This dataset was not processed further.

We were interested to see if the Netbox tool could be used to identify patient specific modules, according to patient prognosis scores, and then look at module conservation across all samples. Therefore, CAM vs. ANM good and CAM vs. ANM bad differentially regulated genes were analysed to reveal common and unique modules. The good patient prognosis network consisted of 80 genes with no statistically enriched linkers ($p \leq 0.05$), which resulted in 15 modules (8 modules >4 gene members) (Figure 4.14A). The bad patient prognosis network consisted of 72 genes with 36 linkers ($p \leq 0.05$), resulting in 10 modules (8 modules >4 gene members) (Figure 4.14B).

There are 11 genes that are commonly mapped to modules within both the 'good' and 'bad' CAM vs. ANM patient prognosis groups, which are represented by large network nodes (Figure 14A and B). The common differentially regulated genes include; CBL, COX4I1, KIF5B, KPNB1, KRAS, MAP3K2, NRIP1, PRPF4, SF3B1, SHC1 and TPR. This list includes several proto-oncogenes and genes involved in common signalling transduction pathways, such as CBL, KRAS, SHC1 (Thien and Langdon, 2001). A couple of genes are connected to energy metabolism including COX4I1 and NRIP1. COX4I1 is the terminal mitochondrial respiratory chain enzyme, producing ATP from the citric acid cycle, fatty acid metabolism and amino acid oxidation. It is interesting to note that NRIP is down regulated, as RIP proteins are nuclear receptor interacting proteins, which bind to and inhibit nuclear receptors regulating energy metabolism enzymes (Rosell et al., 2011). The remaining genes are involved in nuclear export, pre-mRNA processing and splicing. Notably, all of these genes change in the same direction in good and bad prognosis datasets, with around half displaying progressive trends in bad prognosis patients. These commonly differentially regulated genes, may represent possible 'driver genes', who's change in expression could control a large number of biological processes and be causative of the gastric cancer phenotype.

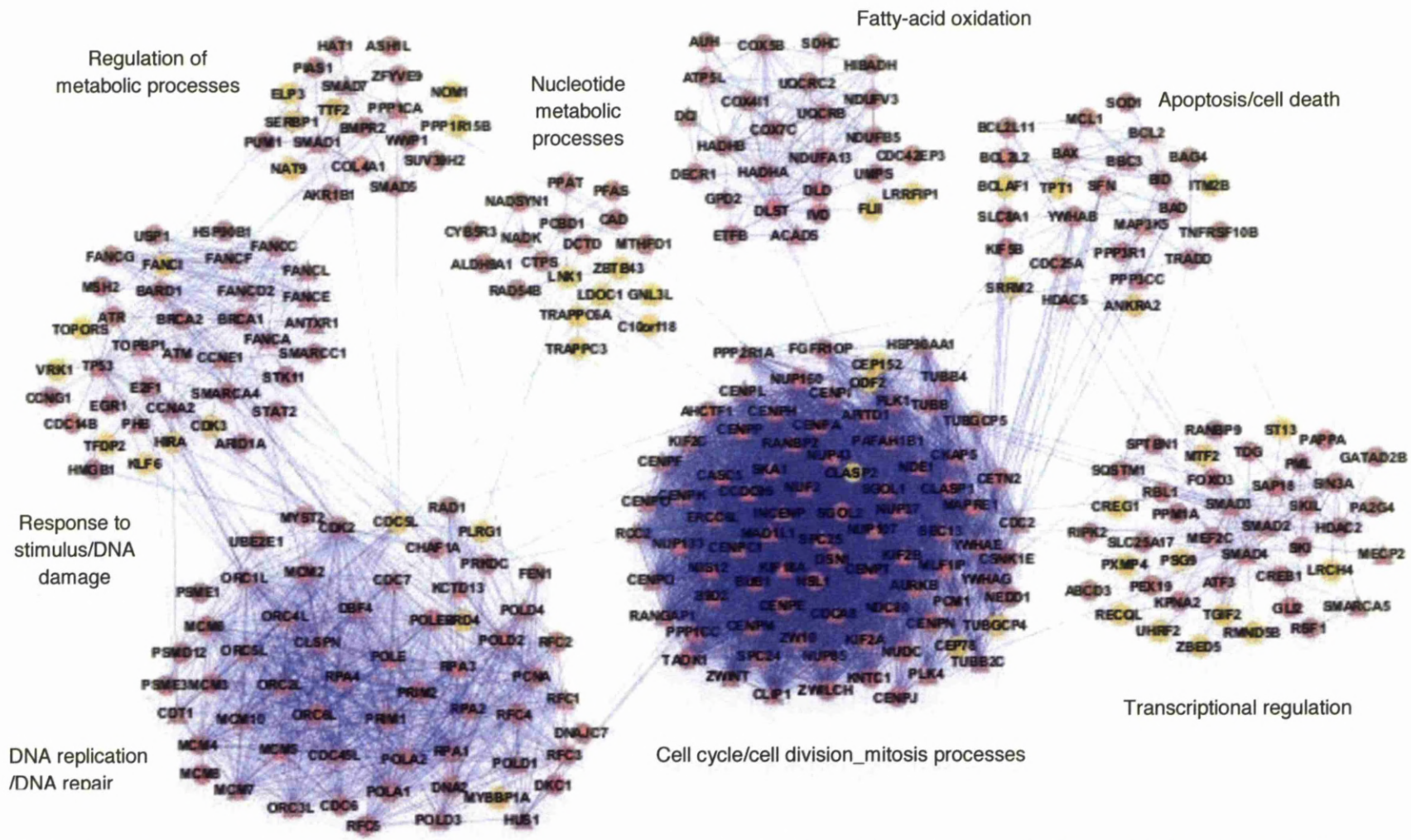



**Figure 4.12.** CAM vs. ANM Netbox Module networks. Modules are densely connected sets of differentially regulated genes (p≤0.05), 8 of the largest modules shown. Differentially regulated genes are shown as circle nodes, yellow nodes represent genes which were previously unable to be mapped to canonical pathways but are identified as module members. Linker genes as shown by triangles and are not differentially expressed but are statistically enriched within the network (p≤0.05). The statistically relevant (p<0.05) biological process relating to modules was defined using the Cytoscape plugin BinGo.

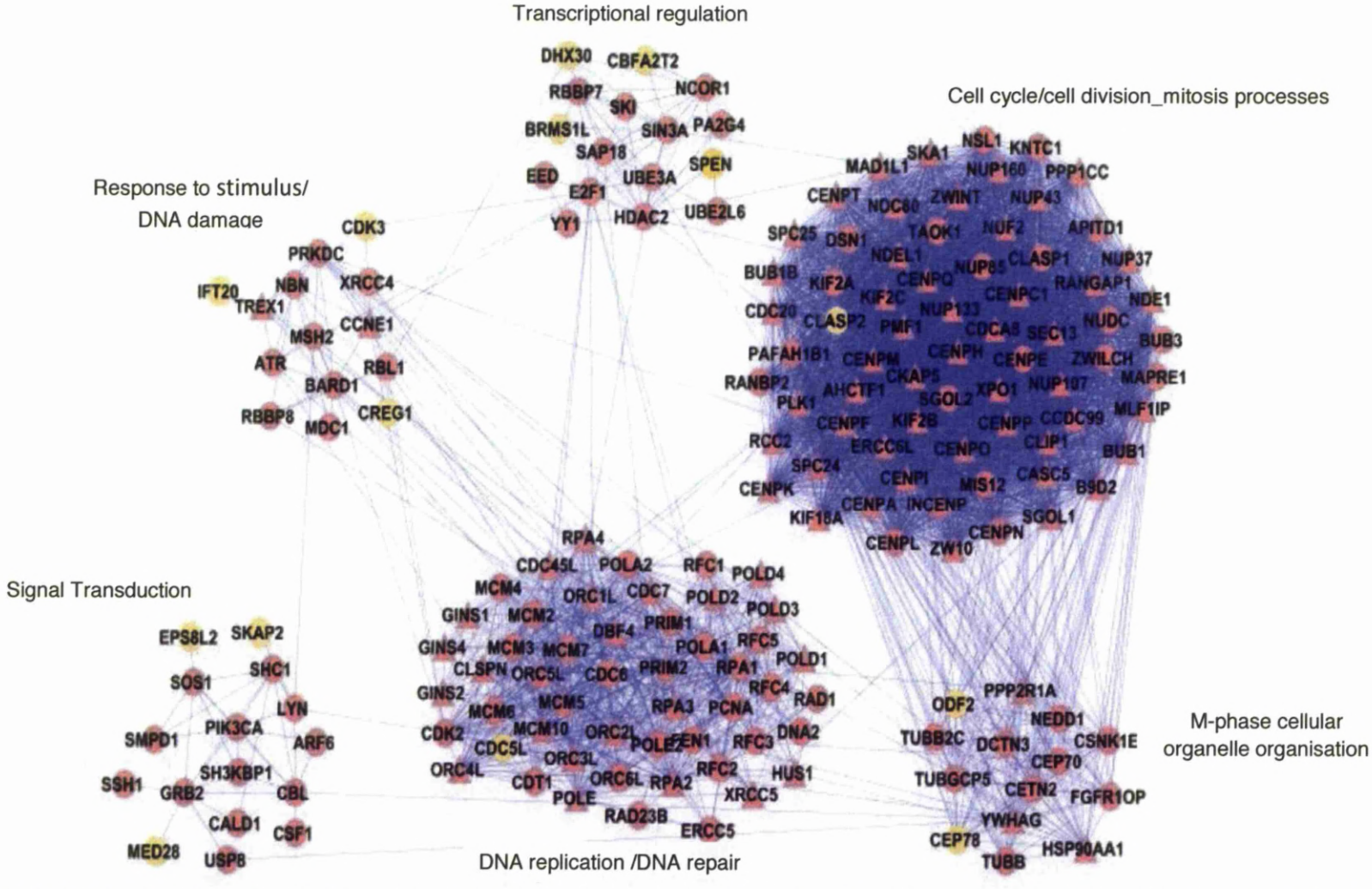



**Figure 4.13.** ATM vs. ANM Netbox Modules. Modules are densely connected sets of differentially regulated genes (p≤0.05), 6 of the largest modules shown. Differentially regulated genes are shown as circle nodes, yellow nodes represent genes which were previously unable to be mapped to canonical pathways but are identified as module members. Linker genes as shown by triangles and are not differentially expressed but are statistically enriched within the network (p≤0.05). The statistically relevant (p<0.05) biological process relating to modules was defined using the Cytoscape plugin BinGo.

As differentially regulated genes differ between 'good' and 'bad' prognosis, common genes will link to different differentially regulated genes, resulting in modules clustering differently and the assigned biological process may changing slightly. Both prognosis groups have similar signalling and energy metabolism modules, the good patient prognosis groups other modules tend to be translation and transcription related, whilst the bad patient prognosis group modules tends to be mitosis, cytoskeleton organisation, chromatin remodelling and gene silencing via mRNA and epigenetics. Larger differentially regulated gene sets, as for the general CAM vs. ANM and ATM vs. ANM datasets, may be required to reduce the amount of variation in module biological process assignment.

The two modules labelled with a blue star (Figure 4.14) represent modules were no statistically relevant biological process could be assigned, these relate to module 1 in the good patient prognosis set and module 0 in the bad patient prognosis module. Comparing modules between the two datasets, the gene expression direction of change is mixed within most modules. Across the good patient prognosis set, genes involved in transcription and regulation of signalling pathways are universally down regulated. Within the energy metabolism/fatty acid oxidation module, all genes involved in are up-regulated in the bad patient prognosis group, and 3/4 of genes are up-regulated in the good patient prognosis group.

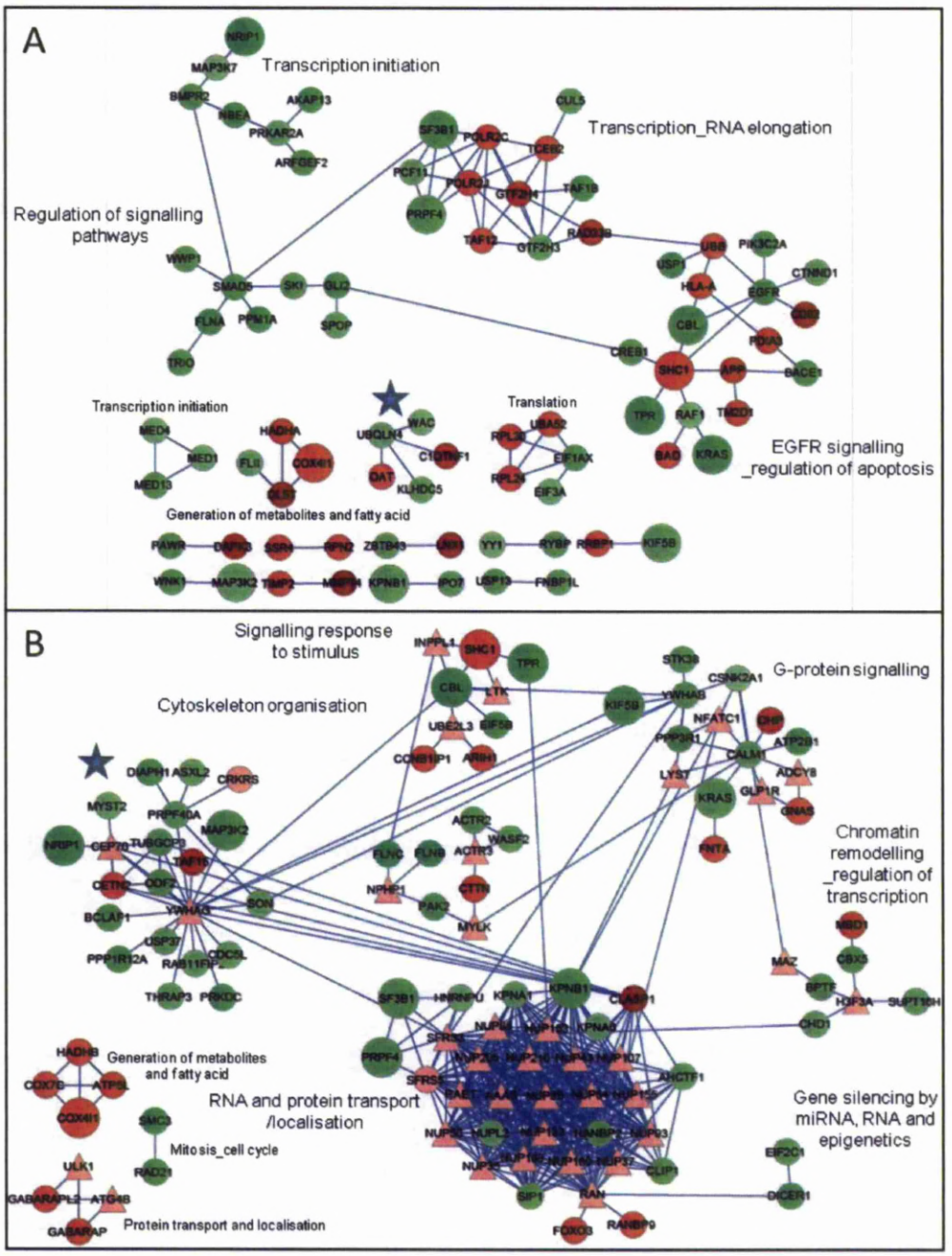



**Figure 4.14.** A) Good Netbox prognosis module network B) Bad Netbox prognosis module network. Modules are densely connected sets of differentially regulated genes ($p \leq 0.05$), as shown by nodes, green representing down-regulated genes and red representing up-regulated genes. Large nodes are common differentially regulated genes within the good and bad patient prognosis datasets. Linker genes as shown by pink triangles are not differentially expressed but are statistically enriched within the network ($p \leq 0.05$). The statistically relevant ($p < 0.05$) biological process relating to modules was defined using the Cytoscape plugin BinGo, blue stars represent modules were no statistically relevant biological process could be assigned.

## 4.3 Discussion

Preliminary network analysis revealed a high level of connectivity between differentially regulated genes within the interactome. The large amount of connectivity visualised between genes, which are members of different pathways demonstrates that although we tend to think of pathways as having linear order, they obviously do not work in isolation. They are highly connected processes that share components; therefore, disruption of shared components can potentially have dramatic effects on multiple pathways. Genes that were unable to be assigned to Metacore, DAVID, Ingenuity or Reactome canonical pathway tools were also highly connected within the interactome and directly interact with canonical pathway members. Therefore, we have demonstrated that integration of pathway enrichment methods with large-scale interactome networks may provide additional data relating to the potential global influence of differentially expressed proteins that link multiple biological processes or proteins that although not currently assigned to defined biological pathways are found to interact with one or more core pathway component. Therefore, may exert influence over one or more differentially regulated pathways. In order to analyse this phenomenon further it is necessary to extract details of all components assigned to each of the differentially regulated pathway so that this gene list can then be imported into an interactome network. However, it is extremely difficult to extract this information for multiple pathways using commercial tools such as Metacore, as the software does not allow large-scale export.

In contrast, the open access database, Reactome does allow pathway information to be freely downloaded. Therefore, with the help of a collaborating statistician we

exported all human biological pathways from the Reactome database using Bioconductor. Hierarchical clustering analysis and multi-dimensional scaling were then used to form groups within the data, and reveal either: pathways that contain similar gene members, or genes that occur in similar pathways. From this analysis we aimed to provide insight into differentially expressed genes that had the potential to exert maximal impact on the system by virtue of involvement in multiple biological processes and the spectrum of processes that may be effected by key differentially regulated genes in cancer derived or adjacent myofibroblasts.

It became apparent that analysing the datasets separately as 'similarities based on genes' and 'similarities based on pathways' was difficult. The limitation with this multi-dimensional scaling technique was the ability to compare clusters resulting from the two types of analysis. Therefore, upon identification of interesting clusters of over-represented pathways, identifying the relating cluster of significantly changed gene was impossible. A conclusion was reached that pathways and genes needed to be plotted simultaneously, using a multidimensional scaling technique called correspondence analysis (Chapter 5, section 1.6). For this reason, dense central clusters of genes and pathways were analysed in detail within this chapter, but details of other over-represented pathways and significantly changed genes are provided within Supplementary Chapter 1. As the purpose of this analysis was to reveal pathways and genes that have the potential to affect a large number of processes, dense clusters were analysed in detail. Although an over-view of potentially important pathways and genes included within all clusters, including those identified within Supplementary Chapter 1 are discussed below.

Primarily, plotting the pathways and genes as interactive 3-dimensional plots dramatically aids the identification of real clusters within the dataset in comparison to previous results obtained by canonical pathway tools (Chapter 3). Applying this approach across all three datasets, revealed that processes involved in DNA and chromosomal maintenance and repair, were over-represented with the overwhelming majority of differentially regulated genes being down regulated. As were pathways and genes involved in DNA synthesis/replication and replication stress. One such re-occurring over-represented process was the Fanconi Anemia pathway, which works to maintain genomic stability. Individuals displaying the autosomal recessive disorder have increased susceptibility to cancer, although the number of cases detected in gastric cancer are low (Pavithran et al., 2002). Again, the overwhelming majority of differentially regulated genes in these pathways were down-regulated. Finally apoptosis appeared increased across all datasets.

In the CAM vs. ANM and CAM vs. ATM datasets, there are clusters of pathways and groups of genes, representing immune related processes. With large proportions of differentially regulated genes up regulated in interferon signalling and interactions between lymphoid and non-lymphoid cells. Interferons are released by host cells upon detection of tumour cells, activating an immune response. Interferons along with several antigen-presenting proteins (HLA's) have been up regulated within these myofibroblasts, suggesting a heightened immune response in cancer associated myofibroblasts.

There are very striking re-occurring clusters of pathways and genes involved in energy metabolism over-represented within all datasets, specifically relating to

mitochondrial β-oxidation pathways. All differentially regulated genes across all datasets are up regulated metabolic enzymes, suggesting an increase in energy generation. Within the CAM vs. ATM dataset, the related pathway 'Activated AMPK stimulates fatty-acid oxidation' is over-represented. Activation of AMPK stimulates fatty acid oxidation and inhibits fatty acid synthesis; genes involved in this pathway are both up and down regulated. Also in the CAM vs. ATM dataset, enzymes involved in cholesterol biosynthesis are down-regulated, suggesting the end product of fatty acid β-oxidation, acetyl co-A, enters the mitochondria for the citric acid cycle rather than for the biosynthesis of cholesterol. Within the ATM dataset an additional pathway, specifically involved in the metabolism of carbohydrates 'the pentose phosphate pathway' is over-represented.

At first glance, I was surprised with the large proportion of metabolic pathways, with such strong up-regulation of differentially regulated gene members. This pattern of changes would be consistent with the basic principles described in the 'Reverse Warburg effect'. Warburg initially introduced the idea of cancer cells using aerobic glycolysis. He suggested that in cancer cells, cellular respiration is irreversibly damaged, which leads to an increase of energy production by fermentation to replace the lost energy (Warburg, 1956). It has since been proven that the mitochondrial function within cancer cells themselves is not defective (Fantin et al., 2006). That cancer cells use aerobic glycolysis as a selective growth advantage (Heiden et al., 2009) and may be induced by the hypoxia-inducible factor 1-alpha (HIF-1α) (Robey et al., 2005). The 'Reverse Warburg effect', was first suggested in 2009, as extension to Warburg's original observation (Pavlides et al., 2009a). It is suggested that cancer cells induce cells within the cancer stroma to

become activated, to initiate aerobic glycolysis in neighbouring fibroblasts, causing them to secrete lactate and pyruvate in order to 'feed' neighbouring cancer cells. Cancer cells take up these high energy metabolites (pyruvate and lactate) and feed them into their own TCA cycle generating large amounts of ATP for cell proliferation. Therefore, it is possible that the 'Reverse Warburg effect' will explain the observed up-regulation of fatty acid synthesis detected in cancer-associated fibroblasts in this study. It is possible that cancer associated myofibroblasts may utilize ketones as a second rate energy source for themselves, and convert the β-oxidation product acetyl-co-A into pyruvate to feed the cancer cells. This issue is discussed further in Chapter 5.

In addition, the 'Reverse Warburg effect' would also explain the observed increase in the expression of trans-membrane energy metabolite transporters, which was detected in all three datasets. Specifically, in the CAM vs. ANM dataset, which has the largest number of up regulated transporters, including the differentially regulated proton-coupled monocarboxylate transporter, solute carrier family 16 member 3. Monocarboxylate transporters are essential for the transport of pyruvate and lactate across the cell membrane (Morris and Felmlee, 2008). In addition, two facilitated glucose transporters were also found to be over-expressed; solute carrier family 2 member 10 and solute carrier family 2 member 11. In cancer cells, high levels of cell proliferation requires large amounts of energy, therefore glucose is actively imported into the cell and glucose transporters are often up-regulated (Macheda et al., 2005). Within the other two datasets, only SLC16A3 is also over-expressed in the CAM vs. ATM dataset, and SLC2A11 is over-represented in the ATM vs. ANM dataset.

A further trend, which appeared in all datasets was the fact that signalling pathways seem to be most centrally located within multidimensional scaling plots. With a number of signalling cascade genes, displaying bottleneck properties, playing roles in very large numbers (20+) of signalling pathways. As these genes play roles in large numbers of over-represented pathways, they are potential therapeutic targets, if the processes they control appear causative of the cancer phenotype. The majority of signalling pathways are down regulated in all three datasets. Other main cross-talking genes across all datasets include several types of integrin, ubiquitin, and ribosomal protein subunits. The set of over-represented pathways involved in fatty acid metabolism and metabolite transport although striking, they are by far not the most connected interconnected pathways, mitochondrial β-oxidation enzymes show crosstalk between a maximum of seven over-represented metabolic pathways.

We wanted to deduce using Netbox (Cerami et al., 2010) whether our networks of differentially regulated genes were modular. Identification of modules within networks provides invaluable information about the topology of the dataset. It has been suggested that functional modules are a key level of cellular organisation, with functional module separation being due to their molecular/chemical specificity or cellular localisation. The very structure of modules, gives them very distinct characteristics. Changes within a module would be possible without affecting the entire network architecture. Similarly, changes in connections between modules enable un-related functions to affect one another. It is suggested that modular structures may facilitate evolution, as if molecules where un-structured or not modular then it would be difficult to change a gene as it would affect all proteins in

the cell, the positive alteration of the changed gene may be a disadvantage to others. The advantage of modules in evolution is that changing a gene within a module may predominantly affect other genes within that module, or neighbouring modules which are connected by linker genes. This therefore makes it easier to modify, add or delete genes, changing the way in which modules evolve and adapt to new changes (Hartwell et al., 1999). Identification of modules within our datasets may provide greater insight into of the genes affected upon differential regulation of a specific gene, and the biological processes directly linked. Importantly, clusters of modules in the CAM vs. ANM and ATM vs. ANM datasets reveal real biological processes, demonstrating that modules detected within the network are not just an artefact of biological networks, as suggested previously (Hallinan, 2004).

In glioblastoma the p53 pathway is altered but in the different cancers, different members controlling this pathway are altered (Cerami et al., 2010). Therefore, it has been stated that modules are universal but the genetics within and controlling the modules may vary, and this is exactly what we found. We have identified key modules occurring within both the CAM and ATM myofibroblasts, and highlighted the different gene members altered, based on the severity of the cancer.

The CAM vs. ANM and ATM vs. ANM datasets generated high-scaled network modularity scores, demonstrating more local modularity than would be expected by random chance and were able to incorporate redundant canonical pathway genes into biological modules. The biological processes common within CAM and ATM myofibroblasts included: response to stimulus/DNA damage; DNA replication/DNA repair; transcriptional regulation; cell cycle/cell division mitosis and M-phase

cellular organelle organisation. The cancer unique modules include; regulation of metabolic processes, nucleotide metabolic processes, fatty- acid oxidation/energy metabolism and apoptosis. Whereas, there was only one unique module within ATMs; signal transduction. It is very interesting that metabolic modules are unique within cancer myofibroblasts, due to the earlier identification of genes related to the Warburg effect, suggesting this process may define CAMs from ATMs and give them the ability to support and promote tumour progression. Although ATMs displayed a larger number of differentially regulated genes than corresponding cancer derived myofibroblasts Netbox analysis showed that CAMs actually have a different network topology with a wider range of highly connected modules than observed in ATMs.

In addition, Netbox was also used to identify modules relating to particular patient prognosis scores. When comparing 'good' and 'bad' patient datasets, genes that are differentially regulated in both 'good' and 'bad' cancer sets appear to fall within similar (but not identical) modules. Both prognosis groups have similar signalling and energy metabolism modules, however, good prognosis patient groups also have other modules relating to translation and transcription, whilst the bad patient prognosis group also have modules relating to mitosis, cytoskeletal organisation, chromatin remodelling and gene silencing via mRNA and epigenetics.

However, although there are similarities between the 'good' and 'bad' prognosis datasets, the number and type of genes that are differentially regulated within each module clearly differ. Several genes are commonly mapped including several proto-oncogenes, genes involved in common signalling transduction pathways, and a

couple of genes related to energy metabolism. Notably, all of these genes change in the same direction in good and bad prognosis datasets, with around half displaying progressive trends in bad prognosis patients. These common differentially regulated genes, may represent possible 'driver genes', who's change in expression could control a large number of biological processes and be causative of the cancer phenotype.

# Chapter Five: Refined analysis of comparative gene expression profiles

## 5 Introduction

With the need for personalised medicine in the future, patient prognosis scores have been used to try to predict the stage and type of cancer, in order to administer the appropriate treatment. Previously we have used individual patient prognosis scores to determine which patients fall within 'good' and 'bad' patient prognosis groups. Recently we have been supplied with additional patient survival scores. The appropriateness of these survival scores to be able to predict patient prognosis needs to be assessed. Within this chapter, two different types of analysis will be carried out. Firstly, to compare the molecular and cellular characteristics of cancer myofibroblast to adjacent myofibroblasts, in order to understand the biological processes and transcription factors which potentially drive tumour related phenotypes. As myofibroblasts derived from the site of tumours have been shown to grow faster, with increased angiogenic potential and resistance to therapeutic drugs (Barclay et al., 2005). This process of non-cancerous cells conditioning was initially referred to as 'field cancerisation' (Slaughter et al., 1953).

In this study isolated populations of cancer associated myofibroblasts (CAM), matched adjacent (ATM) and absolute normal (ATM) myofibroblasts were used to compare the molecular changes that occur in myofibroblasts derived from patients with early ('good') or late stage ('bad') gastric tumours.

## 5.1 Data analysis

### 5.1.1 Patient survival and prognosis

Analysis of patient survival and histopathology data (Table 5.1) shows that in several cases the patient prognosis scores do not directly correlate with patient survival period. In this study prognosis scores < 9 were classified as less advanced tumours, representing the 'good' patient sub-group, whilst scores >9 represent advanced or metastatic tumours, representing the 'bad' patient sub-group. In terms of survival data, patients who survived longer than 24 months were initially classified as having 'good' survival, whilst those patients who lived less than 24 months were classed as having 'bad' survival. However, this criterion was ultimately refined to reflect positive or negative lymph node status. Significantly this did not change the allocation of patients into 'good' or 'bad' sub-groups. Although patients with more advanced tumours may be expected to live for shorter periods this was not always the case as patients 4, 5, 9 and 13, (samples sz192, sz194, sz271 and sz187 respectively) have conflicting prognosis and survival scores.

Patient 9 (sz271) and patient 13 (sz187) have very 'bad' histopathology scores, yet are still alive, current survival scores are >34 months and >43 months respectively. In addition, although patient 4 (sz192) and patient 5 (sz194) do not have conflicting histopathology and survival scores, they are both borderline between 'good' and 'bad' patient sub-groupings. For example, Patient 4 has a relatively 'bad' histopathology score (11), yet has lived for 22 months, whilst patient 5 has a very 'good' histopathology score (4), yet died after 25 months.

| Patient Label | Sample | Age | Gender | Location of Tumor | Lauren Classification | PROGNOSIS SCORE | 'good' or 'bad' Sub-groups | Survival (Months) | 'good' or 'bad' Survival |
|---|---|---|---|---|---|---|---|---|---|
| 1 | sz42 | 72 | M | antrum corpus border | medullar (non-Lauren) | 5 | G | >59 | G |
| 2 | sz45 | 82 | M | antrum | intestinal | 11 | B | 3 | B |
| 3 | sz190 | 66 | F | antrum corpus border | mixed | 13 | B | 5 | B |
| 4 | sz192 | 50 | F | antrum corpus border | diffuse | 11 | conflicting | 22 | conflicting |
| 5 | sz194 | 76 | M | antrum | intestinal | 4 | conflicting | 25 | conflicting |
| 7 | sz198 | 77 | M | antrum | intestinal | 7 | G | >34 | G |
| 8 | sz268 | 76 | M | antrum | intestinal | 11 | B | 15 | B |
| 9 | sz271 | 72 | M | corpus | mixed | 12 | conflicting | >34 | conflicting |
| 10 | sz294 | 84 | F | antrum corpus border | intestinal | 7 | G | >31 | G |
| 11 | sz305 | 59 | F | antrum and corpus | diffuse | 12 | B | 17 | B |
| 12 | sz308 | 51 | M | antrum corpus border | mixed | 12 | B | 9 | B |
| 15 | sz389 | 67 | M | antrum | intestinal | 8 | G | >31 | G |
| 13 | sz187 | 39 | F | antrum corpus border | intestinal | 11 | conflicting | >43 | conflicting |
| 14 | sz197 | 54 | M | antrum corpus border | diffuse | 6 | G | >45 | G |

**Table 5.1**. Patient sample information, including new survival data. Patient prognosis scores ≤ 9, represent 'good' patient sub-groups. Whilst those patient prognosis scores >9 represent 'bad' patient sub-groups. Survival scores ≤ 24 months represent patients with 'bad' patient survival, whilst those with survival scores > 24 months represent 'good' patient survival. Conflicting prognosis score and survival scores are stated as such.

Patient prognosis groups were reclassified, in light of newly available patient records (Table 5.2). The Patient 14 sample (sz197) was not included in the 'good' patient sub-group, as no matched adjacent sample was available for this patient. To detect changes in the 'good' cancer subset vs. absolute normal and the 'bad' cancer sub-set vs. absolute normal datasets, the data was processed as explained in methods chapter 2, section 2.1.1.1; In brief: background oligonucleotide IDs are detected and statistically significant oligonucleotide IDs were defined using un-paired T-Tests and p-values corrected using the Benjamini Hochberg FDR calculation. Finally, oligonucleotide probe IDs were converted into Entrez gene IDs using Metacore™.

| 'Good' patient prognosis groups | |
|---|---|
| 1-CAM | |
| 7-CAM | |
| 10-CAM | |
| 15-CAM | All Absolute normal samples |
| 'Bad' patient prognosis group | |
| 2-CAM | |
| 3-CAM | |
| 8-CAM | |
| 11-CAM | |
| 12-CAM | All Absolute normal samples |

**Table 5.2.** New revised, 'good' and 'bad' patient prognosis groups.

As work on this project progressed further patient information became available (Table 5.3), including data on primary tumour invasion, lymph node metastasis and distant metastasis. It appeared from this data that lymph node metastasis was a 'good' predictive measure of cancer severity and relates well to patient survival scores (C.Holmberg et al, 2012). The utility of this method to distinguish 'good' and 'bad' patient sub-groups was further assessed in this study.

A

| Patient Label | Tumour Staging |
|---|---|
| 1 | pT1N0M0 |
| 2 | pT4N2M0 |
| 3 | pT4N4M1 |
| 4 | pT3N1M0 |
| 5 | pT1N0M0 |
| 7 | pT2aN0M0 |
| 8 | pT4N2M0 |
| 9 | pT3N1M0 |
| 10 | pT3N0M0 |
| 11 | pT3N2M0 |
| 12 | pT1N3M0 |
| 15 | pT3N1M0 |
| 13 | pT4N2M0 |
| 14 | pT2N0M0 |

B

| | |
|---|---|
| Primary Tumour (T): | T1: Tumour invades lamina propria (T1a) or submucosa (T1b)<br><br>T2: Tumour invades muscularis propria (T21a) or subserosa (T2b)<br><br>T3: Tumour penetrates serosa (visceral peritoneum) without invasion of adjacent structures<br><br>T4: Tumour invades adjacent structures |
| Regional Lymph Nodes (N) | N0: No regional lymph node metastasis<br>N1: Metastasis in 1 to 6 regional lymph nodes<br>N2: Metastasis in 7 to 15 regional lymph nodes<br>N3: Metastasis in more than 15 regional lymph nodes |
| Distant metastasis (M) | M0: No distant metastasis<br>M1: Distant metastasis |

**Table 5.3**. Primary tumour, regional lymph node metastasis and distant metastasis information. A) Scores relating to individual patients B) Key providing description of primary tumour, regional lymph node and distant metastasis scores assigned.

### 5.1.2 Refined Principal Component Analysis

PCA analysis of CAM vs. ANM patient samples B ł ATM vs. ANM samples, before and after batch correction (also discussed in Chapter 3) are shown in Figure 5.1 shows. This analysis shows that global gene expression profiles for CAMs and ATMs are distinct from ANM gene expression profiles. Although global gene expression profiles of cancer and adjacent samples are very similar (Figure 5.2A&B), patient pairs correction does improve separation of the two sample types (Figure 5.2C&D).

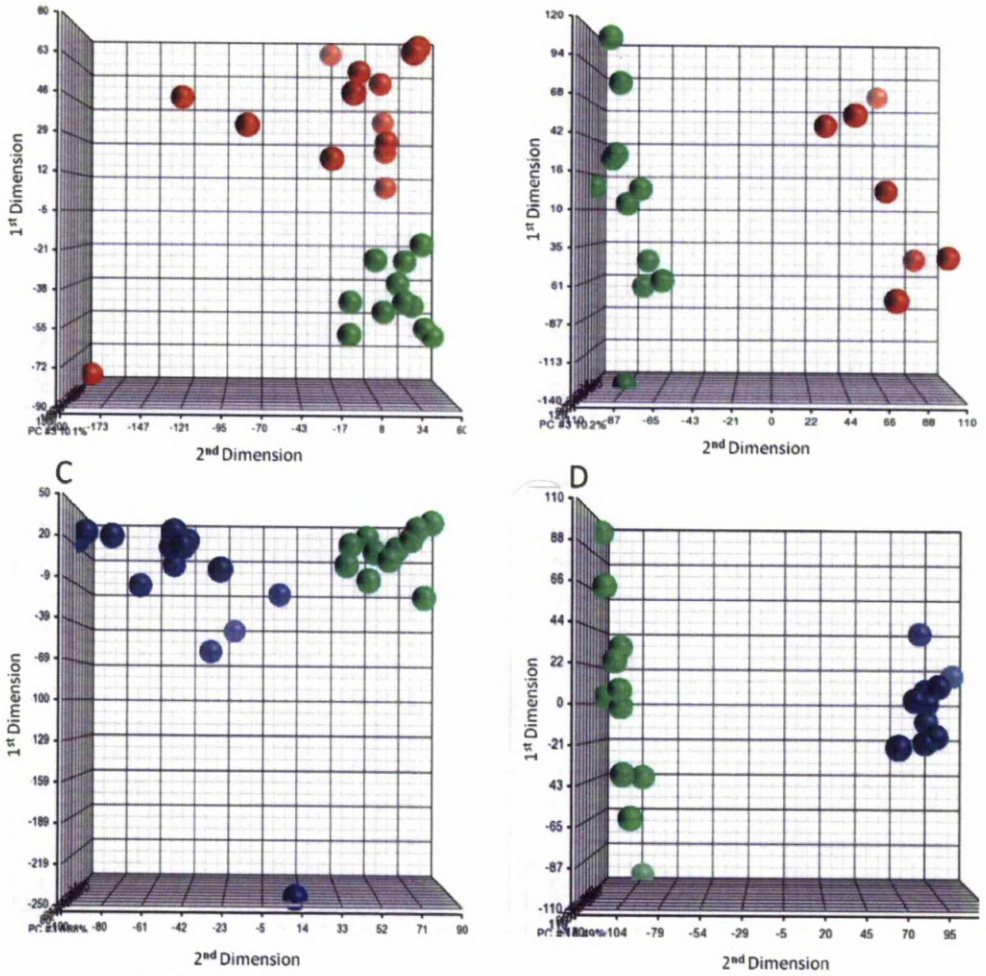



**Figure 5.1.** Principal component analysis. A) Un-corrected cancer (red) and absolute normal (green) patient samples. B) Batch corrected cancer and absolute normal samples C) Un-corrected adjacent (blue) and absolute normal (green) samples D) Batch corrected adjacent and absolute normal samples.

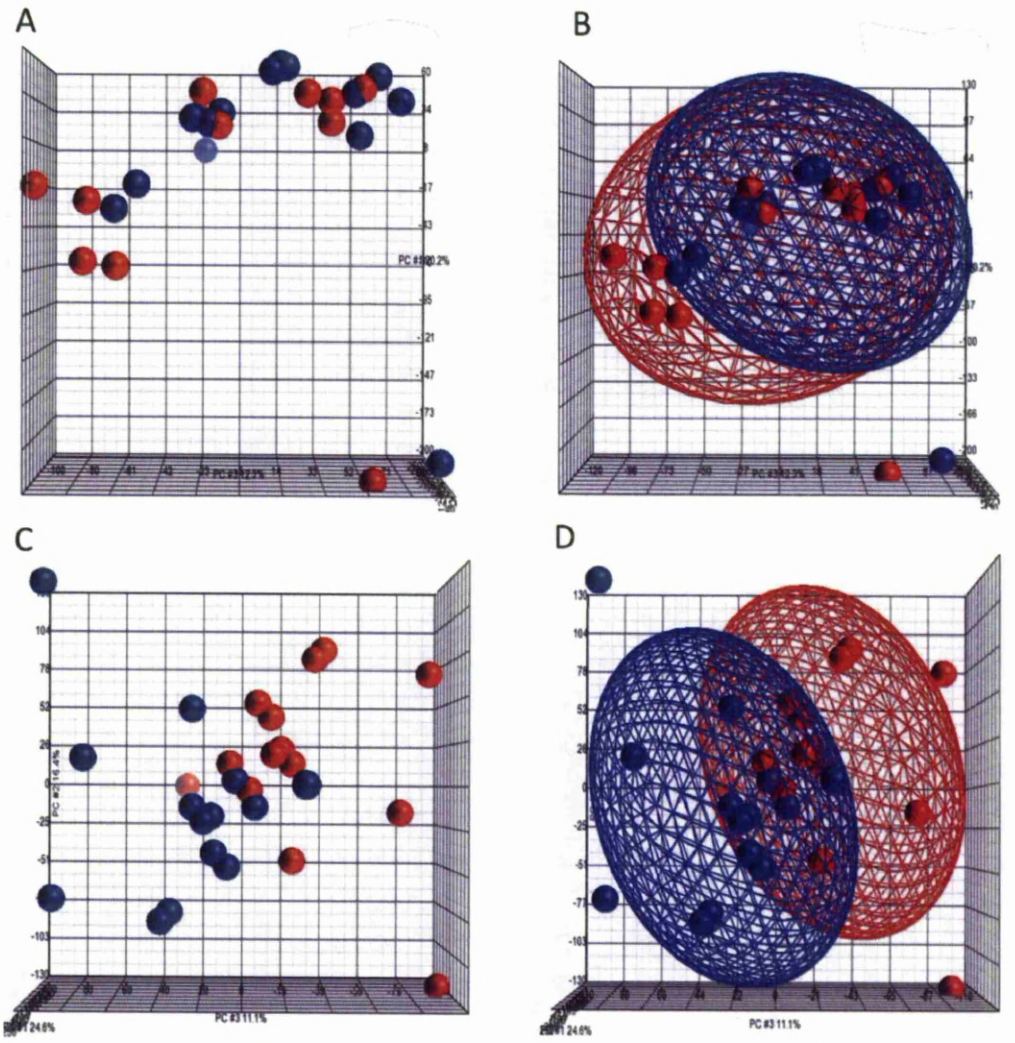



**Figure 5.2**. Principal component analysis of CAM (red) and ATM (blue) patient samples. Panels A and B show un-corrected data while panels C and D show patient-pair corrected data. Cages represent ellipsoids, which depicts the spread of the data based on the centre of origin.

Lists of statistically significantly changed genes from each dataset were generated in Parteks® agglomerative hierarchical clustering method with Euclidean distance and average linkage. The hierarchal clustering analysis was used to cluster patients based on the similarity of their gene expression profiles. This clustering analysis was carried out using varying fold-change thresholds to generate hierarchical cluster

plots to determine the minimum fold-change threshold that effectively segregated myofibroblast subgroups, whilst retaining the highest number of genes possible. As shown in Figure 5.3, the samples within the CAM vs. ANM and ATM vs. ANM datasets correctly segregate into two distinct groups at any fold-change cut-off between 1 and 2. Thus demonstrating that a p-value significance threshold of ≤0.05 alone, without an imposed fold-change cut-off is sufficient to define the expression profiles that differentiate CAM or ATM samples from ANM samples. All the samples used in this study were from Hungarian patients, apart from one English patient (28AN/AV53). Interestingly, within the ATM vs. ANM dataset, the English patient sample was the first to un-cluster at a 2.2 fold change threshold (data not shown).

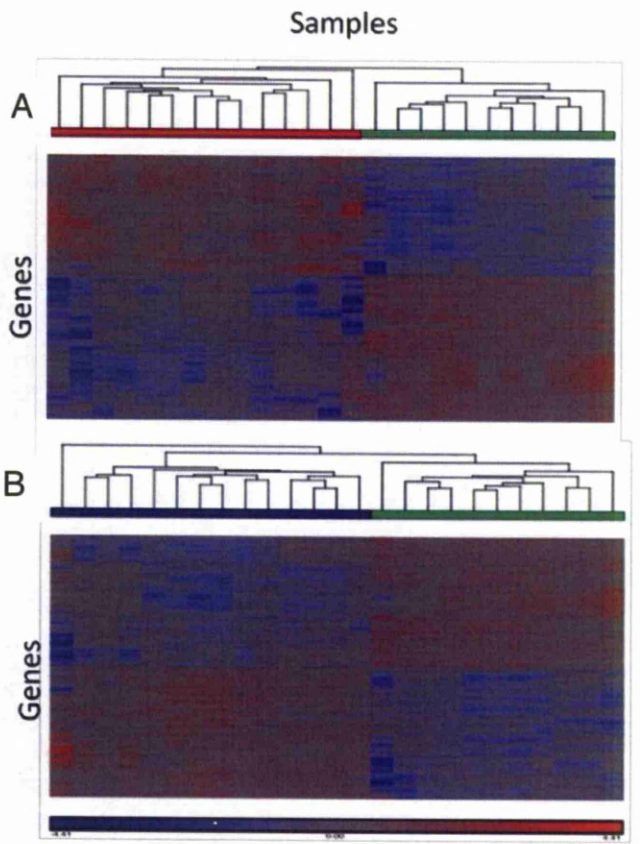



**Figure 5.3** Hierarchal clustering analysis of differentially regulated genes (p≤0.05, no fold change cut-off). (A) CAM vs. ANM samples, (B) ATM vs. ANM samples. Trees represent the similarity between individual CAM (red), ATM (blue) and ANM samples (green).

Data for the CAM/ATM comparison does not separate well on p values alone (Figure 5.4A) for the CAM vs. ATM dataset. Therefore, for this data set it was important to define the optimal fold-change cut-off to apply, whilst retaining the maximum number of genes. Although the samples separate perfectly when applying a 2-fold change cut-off, a 1.6 fold change cut-off also provides separation of most CAM and ATM samples (Figure 5.4 B and C), and this threshold is often selected as an arbitrary cut-off in many microarray studies. To establish an appropriate fold-change cut-off, for use in this study gene lists obtained using a 1.6 or 2 fold-change were used to compare pathway over-representation profiles using Metacore™. Metacore™ pathway analysis revealed that the application of a 2 fold change cut-off resulted in such a small gene list, that mapping genes to over-represented pathways proved difficult, with a maximum number of 3 differentially regulated genes being mapped to the most significantly enriched pathway, therefore a 1.6 fold change cut-off was deemed acceptable for pathway analysis (Section 5.3.3).

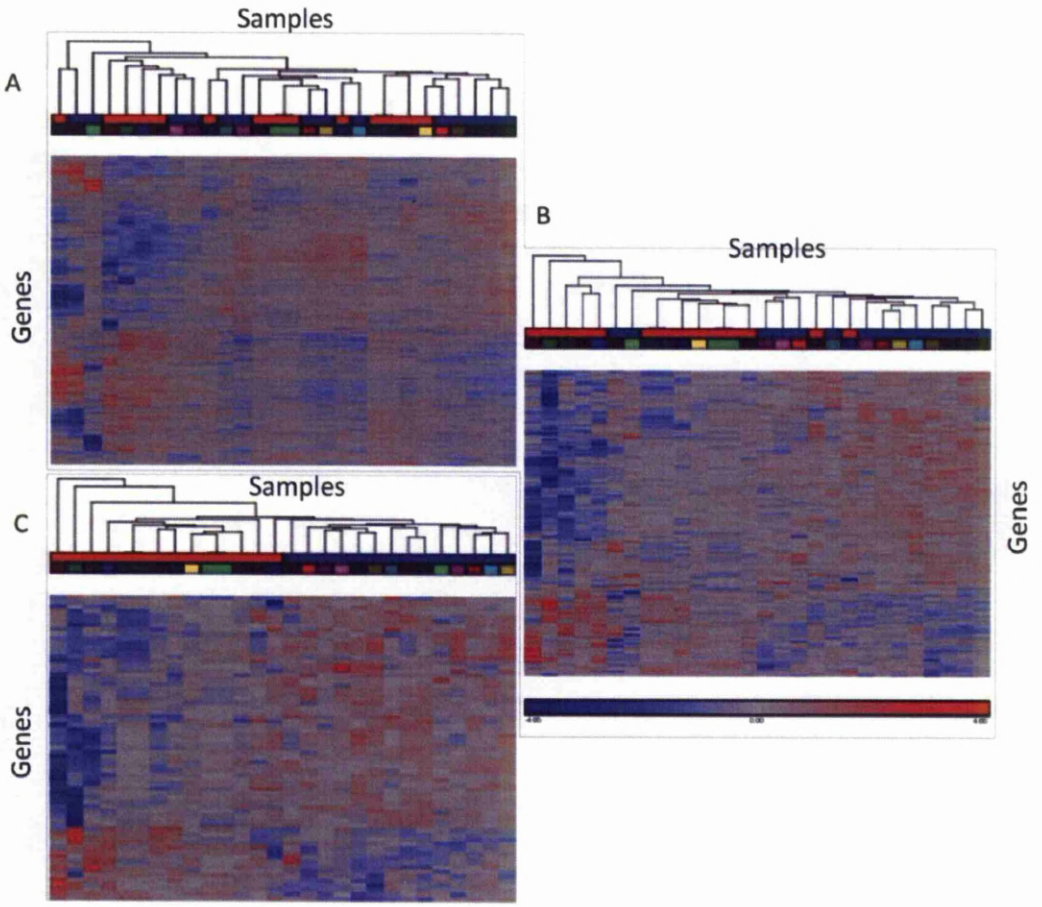



**Figure 5.4.** Hierarchal clustering analysis of differentially regulated genes (p≤0.05) CAM vs. ATM (A) No fold change cut-off. (B) 1.6 fold change cut-off. (C) 2 fold change cut-off. Trees represent similarity between individual CAM (red) and ATM samples (blue).

### 5.1.3 Correlation analysis

#### 5.1.3.1 Relationship between prognosis score and gene expression profiles.

All studies described in this chapter were performed after removal of ATM sample Sz45/22 and ANM sample Sz41/2. All three datasets, CAM vs. ANM, CAM vs. ATM and ATM vs. ANM were analysed to determine which genes exhibit expression trends that correlate with patient prognosis scores (Table 5.1). An example of the type of plot obtained by this method is shown in Figure 5.5A. Genes with scores

approaching 1 or -1 have a high positive or negative correlation between expression level and tumour classification score for each patient sample. The optimal correlation threshold level was independently defined for each dataset by establishing the minimal level of correlation at which 'good' and 'bad' patient subgroups remained fully resolved. The resulting, subsets of differentially expressed genes represent a trend which upon exploration with bigger cohorts, may provide potential insights into stage-specific gene expression signatures, which may in turn indicate the biological processes that are selectively changed in 'good' and 'bad' prognosis cohorts. Expression profiles for genes with positive correlation scores >0.78 with prognosis scores in CAM vs. ANM myofibroblasts were clustered using the complete linkage algorithm, calculating Euclidean distances to generate a hierarchical clustered heat-map (Figure 5.5B). For this analysis, a correlation threshold of 0.78 was found to be the minimal level at which 'good' and 'bad' patient subgroups were resolved, resulting in a set of 9 genes whose expression profiles reflect the stage of tumour progression. In this analysis, no genes were found to exhibit negatively correlation with prognosis scores (data not shown).

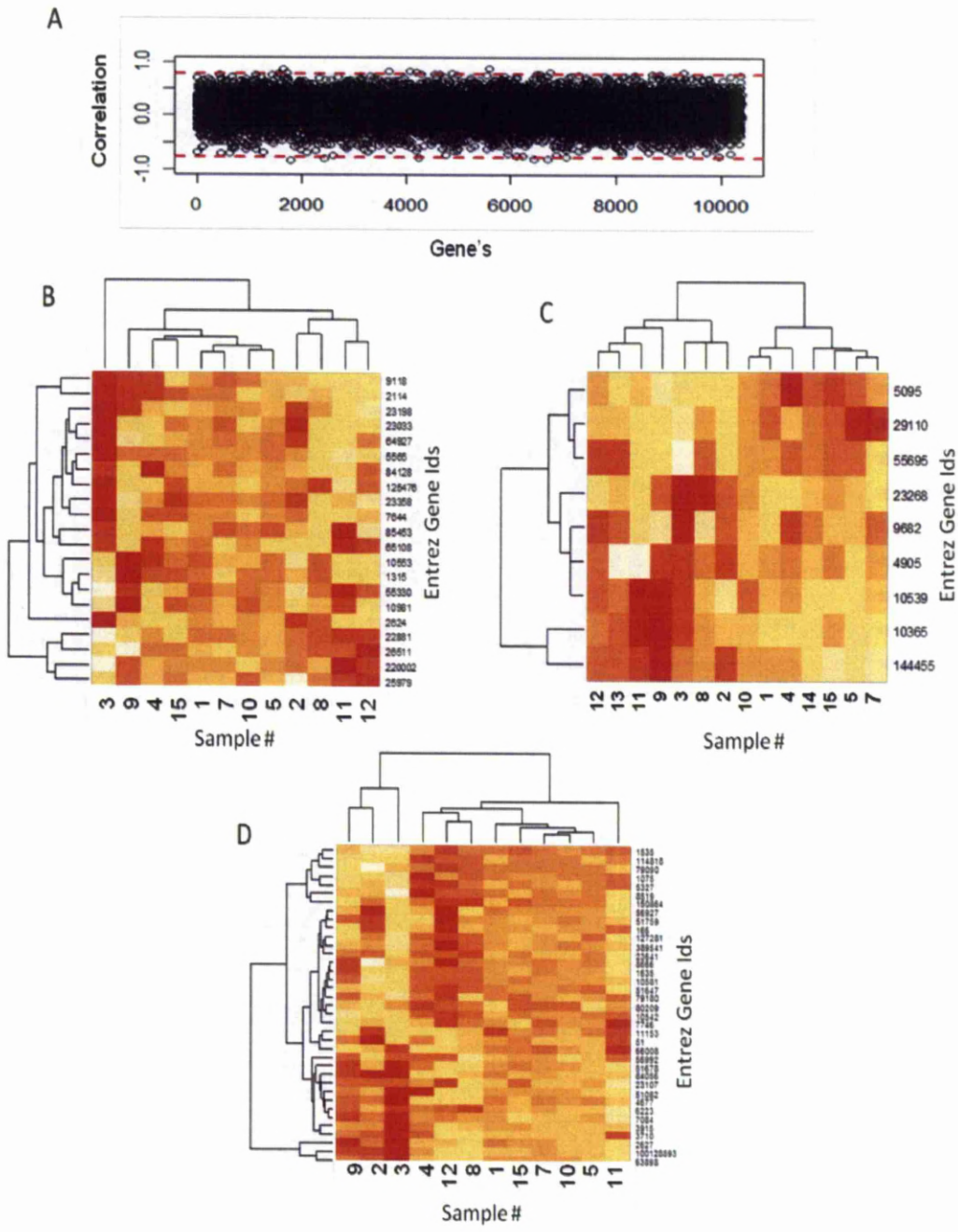



**Figure 5.5**. Spearmans correlation analysis score calculated for each gene. B) ATM vs. AN, genes positively correlated (≥0.77) with prognosis score. C) CAM vs. AN, genes positively correlated (≥0.78) with prognosis. D) CAM vs. ATM, genes correlated with prognosis, ≥0.79. Patient sample numbers are represented on the horizontal axis and Entrez gene Ids are represented on the vertical axis. Yellow represents up regulation. Red represents down regulation.

The nine genes, which show strong correlation with tumour classification scores, indicate trends that may provide insights into tumour development within myofibroblasts derived from the cancer region. Three of these genes have large positive fold changes in patients with prognosis scores >9 ('bad'), compared to patients with prognosis scores <9 ('good') these are: PCCA (Entrez ID 5095) propionyl CoA carboxylase, TBK1 (Entrez ID 29110) TANK-binding kinase 1 and NSUN5 (Entrez id 55695) NOP2/Sun domain family member 5. Whilst the other 6 have large negative fold changes in 'bad' compared to 'good' patient derived myofibroblasts; DNMBP (Entrez ID 9682) dynamin binding protein, KDM4A (Entrez ID 9682) lysine (K)-specific demethylase 4A, NSF (Entrez id 4905) N-ethylmaleimide-sensitive factor, GLRX3 (Entrez ID 10539) glutaredoxin 3, KLF2 (Entrez ID 10365) Kruppel-like factor 2 and E2F7 (Entrez ID 144455) E2F transcription factor 7.

Patient samples 4 and 13 both have relatively 'bad' prognosis scores of 11, however, both were labelled as 'conflicting' (Table 5.1), as these patients both lived longer than may have been expected. Within the correlation analysis, sample 4 clusters with 'good' patients, whereas sample 13 clusters with 'bad' patients. Data from patients' samples 4 and 13 were re-moved and the correlation analysis was repeated, to see if hierarchical clustering of the restricted cohort would generate better segregation of 'good' or 'bad' patient sub-groups. However, removal of these patients failed to improve clustering following correlation analysis (data not shown).

Following correlation analysis of the ATM vs. ANM dataset (Figure 5.5C) a correlation score threshold of 0.77 was found to be optimal. The 'good'/'bad' separation was not as clear as displayed within the CAM vs. ANM dataset but 'bad'

patients displaying two distinct types of 'bad' patient subgroups. No genes showed strong negative correlation with prognosis scores (data not shown).

With respect to CAM vs. ATM correlation analysis, 36 genes that show a strong positive correlation with tumour classification stage score were defined using an optimised correlation threshold of 0.79 (Figure 5.5D). In total 36 genes were found to have larger fold changes in cancer samples from patients with worse prognosis scores (Supplementary table 5.1). The 'good' and 'bad' patient sub-groups separate, with the 'good' patients clustering centrally with two distinct 'bad' patient sub-groups segregating on either side. This separation is much more distinct than the sub-groups identified within the ATM vs. ANM and therefore sub-groups are defined as: 'bad' sub-group A [patients 2(sz45), 3(sz190) and 9(sz271)], and 'bad' sub-group B [patients 4 (sz192), 8(sz265), 11(sz305) and 12(308)]. Interestingly, this data confirms results obtained in the CAM vs. ANM correlation analysis, regarding the clustering of inconclusive patients. In particular, 'inconclusive' patient 4, was now found to be clustered with the 'bad' patient group. As 'inconclusive' patient 13 does not have an associated ATM sample this was not included in the analysis. To understand why such 'bad' patient cohorts may exist, and to try and correctly place sample 4 within a definite myofibroblast subgroup, patients *H.pylori* status was investigated. Interestingly all patients within 'bad' subgroup A were not infected by *H.pylori*, whereas all patients within 'bad' subgroup B were infected, with the majority being the cagA positive strain, which is known to be associated with gastric cancer. Therefore, it appears that comparison to paired ATM samples is needed to correctly define inconclusive patients within 'good' or 'bad' patient cohorts. It is worth noting that no obvious bias in terms of age and sex was

identified in relation to either 'bad' subgroups. Also, neither 'good' nor 'bad' subgroups show any correlation with batch processing dates. Interestingly, all three of the correlation heat-maps show that fold expression changes for genes with high correlation scores in 'bad' patient sub-groups were noticeably larger than those observed for genes showing correlation to prognosis scores in 'good' patient sub-groups.

#### 5.1.3.2 Variance of gene fold changes

To investigate the apparent trends in gene expression observed in correlation analysis heat maps, the variances of genes fold changes for patients within 'good' and 'bad' sub-groups were calculated. Results from this analysis clearly show that 'bad' patients have much larger fold changes in gene expression than 'good' patients (Figure 5.6).

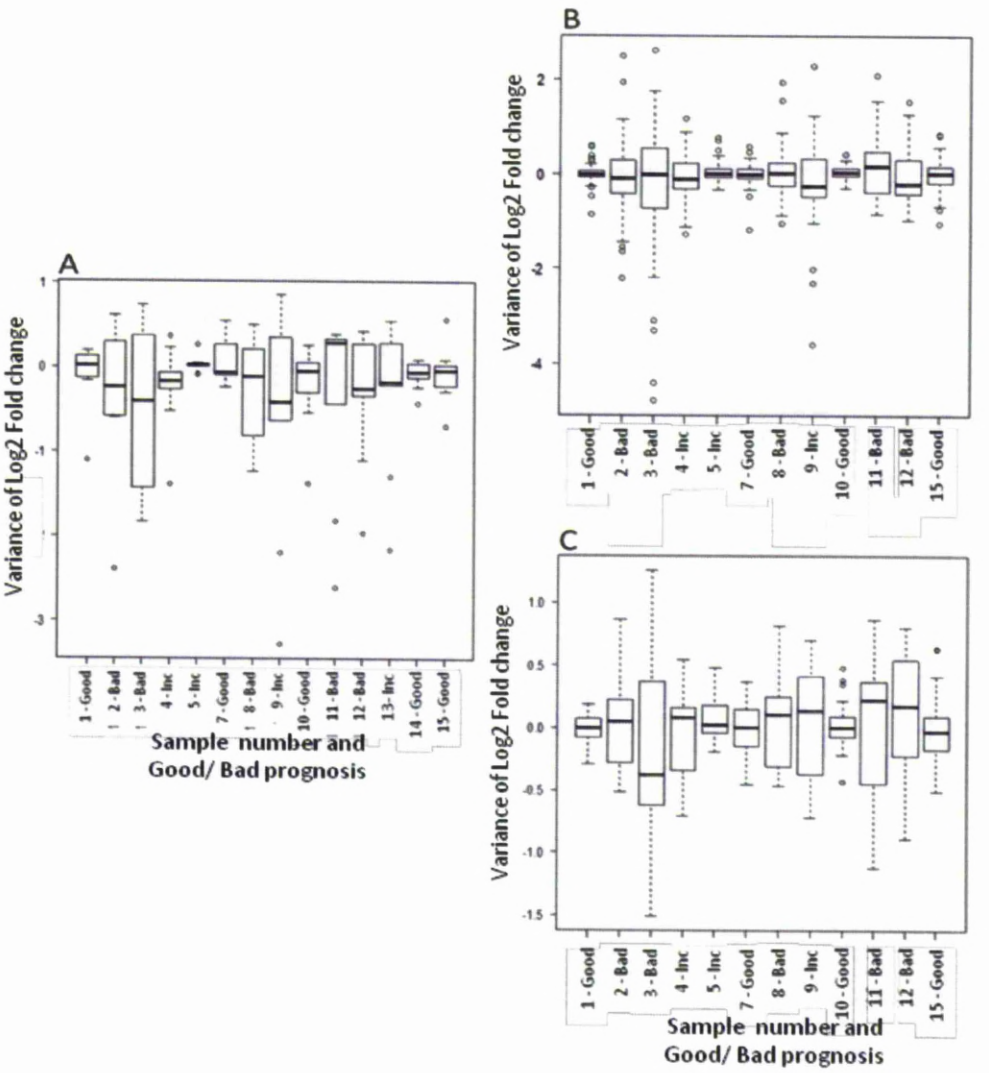



**Figure 5.6**. Box plots representing the variance in Log Fold change of genes for all individual patients. (A) CAM vs. ANM. (B) CAM vs. ATM. (C) ATM vs. ANM. Patient sample Ids and associated patient prognosis category are represented on the x axis, patient prognosis scores<9 are classified as 'Good' and patient prognosis scores >9 are classified as 'Bad'.

As boxplots use log fold changes for all genes, they allow an un-bias way of assessing 'inconclusive' patients. For the case of inconclusive patient 13 ('bad' prognosis score but 'good' survival score), the correlation analysis based on tumour classification score clusters sample 13 with 'bad' patients, in direct contrast to its survival scores (>43 months). Significantly, boxplots of all genes fold-change variances, provide a clear distinction, with sample 13 displaying a large variance.

Since small variances in fold change is associated with 'good' patient tumour classification scores, while large variances in fold change is associated with 'bad' patient tumour classification scores. It is clear that patient samples can be accurately assigned to either 'good' or 'bad' prognosis score sub-groups simply by defining the relative degree of variance in gene expression levels, where 'bad' patient sub-groups exhibit large variance in gene expression profiles while 'good' patient sub-groups show significantly less variance in gene expression.

Inconclusive sample 4 (prognosis score 11 and survival score 22 months) was classified within the correlation analysis as 'good' upon comparison to absolute normal samples and 'bad' when compared to adjacent samples. Boxplots clarify this, as fold-change variances, are small within the CAM vs. ANM dataset and large within the CAM vs. ATM dataset. Placing this patient in alternative prognosis groups depending on the comparative dataset. As stated earlier, upon relation to patients *H.pylori* status patient 4 is infected with the cagA positive *H.Pyori* strain along with all other patients identified within 'Bad' patient prognosis group B. Thus suggesting that the comparison of cancer samples to adjacent samples may be a more accurate method of predicting patient prognosis.

Survival scores generated may not be accurate enough to use alongside gene expression data to predict prognosis due to other factors influencing the longevity of the patients, such as the ability of an individual's immune system to impair growth of cancer cells (Bhardwaj, 2007). Further to the earlier statement regarding patients with 'bad' prognosis scores having larger fold-changes in gene expression patterns. When log fold changes genes are plotted against prognosis scores for each

patient, for the CAM vs. ANM and the CAM vs. ATM datasets (Figure 5.7), patient samples with worse prognosis scores can be seen to have larger variance of all gene log fold changes. In addition, it is clear that patients with worse prognosis scores have a larger variance of log fold changes of the highly correlated genes. By representing the data in this format it allows genes that have large fold changes and highly correlation with prognosis scores to be identified (represented by dashed circles in Figure 5.7B).

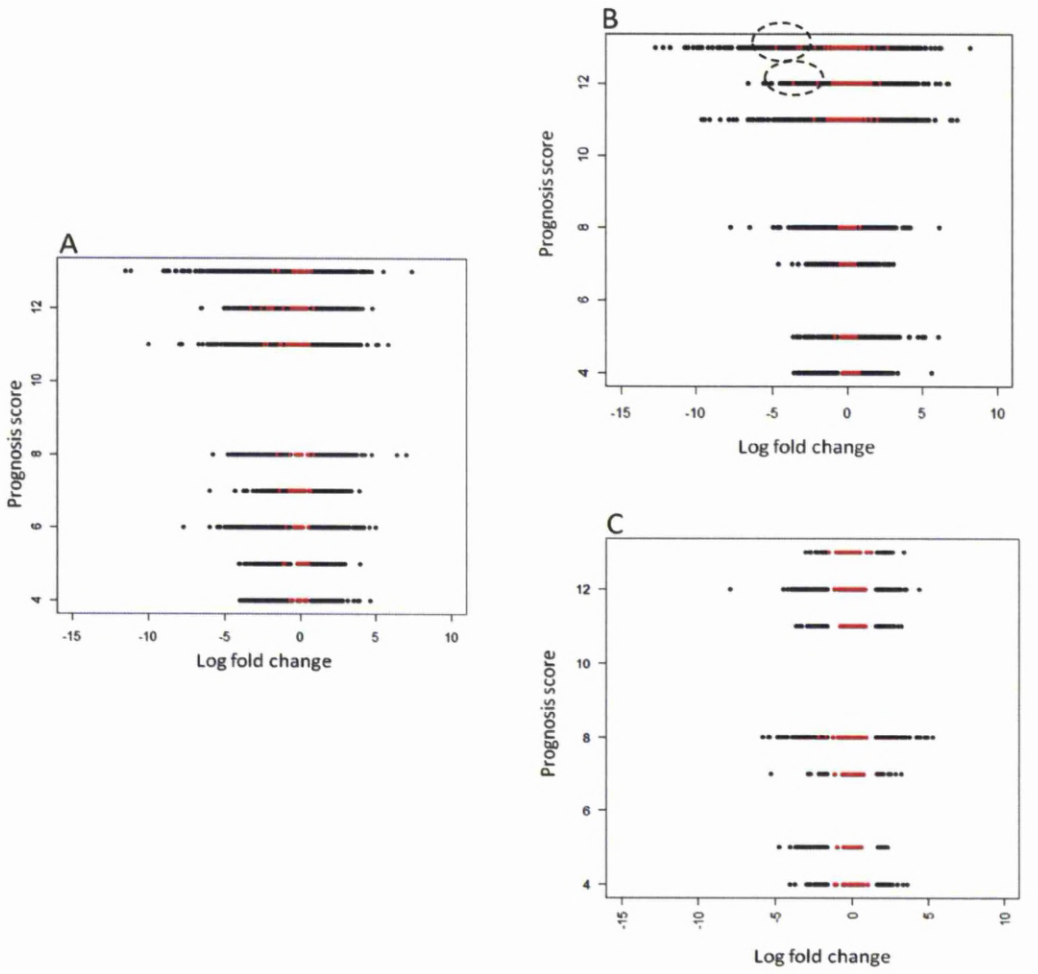



**Figure 5.7**. Individual genes log fold changes plotted against patient prognosis scores. (A) CAM vs. ANM, red points represent genes with optimum correlation scores ≥0.78. (B) CAM vs. ATM, red points represent genes with optimum correlation scores ≥0.77. (C) ATM vs. ANM, red points represent genes with optimum correlation scores ≥0.77. Genes with fold changes < 1.6 are not displayed.

Subsequently, for the CAM vs. ANM and the CAM vs. ATM datasets, genes that show high correlation to tumour classification score and have log fold changes >1.6 or a 2-fold change were identified. Genes with fold changes above a 2-fold change have a better potential of being confirmed using qPCR. Three genes are identified in the CAM vs. ANM dataset as being highly correlated (0.78) and having greater than 2 fold changes (Table 5.4). E2F7 is shown to be highly down regulated in 10/14 patients, having small fold changes (<2 fold) in 4 patients with the best prognosis scores and is increasingly down regulated in patients with poor tumour classification scores. Patient 9 has a tumour classification score of 12 and E2F7 is most highly down regulated in this patient (by 9.6 fold). E2F7 is a transcription factor, specifically a transcriptional repressor that is believed to have negative effects on the regulation of the cell cycle and cellular proliferation (de et al., 2003). The kruppel-like factor KLF2 is a transcription factor down regulated in five patients with the worst prognosis scores. KLF2 is down regulated in many cancers and is seen as a potential tumour suppressor, due to its roles in inhibiting proliferation and migration, and initiating apoptosis and senescence (Black et al., 2001). One particular KLF2 is known to be epigenetically silenced in cancer cells by EZH2 (Taniguchi et al., 2011). DNMBP is a scaffold protein that links dynamin with actin and is thought to be involved in membrane trafficking. Interestingly, DNMBP is only down regulated by >2 fold in CAMs derived from patient 3, which is the patient with the worst tumour classification score of 13, showing >2 fold changes in all three genes (KLF2, DNMBP and E2F7). Mitochondrial enzyme Propionyl-CoA carboxylase (PCCA) was the only highly correlated gene that was added when the fold change

threshold was lowered to 1.6 PCCA was over-expressed in two patients with the worst tumour classification scores, 1.8 fold in patient sample 3 and 1.66 fold in patient sample 9.

| Patient | Prognosis | KLF2 | DNMBP | E2F7 |
|---|---|---|---|---|
| Sample 1 | Prog5 'good' | Entrez | Entrez | -2.14711Entrez |
| Sample 2 | Prog11 'bad' | | | -5.22738 |
| Sample 3 | Prog13 'bad' | -2.6884 | -2.82622 | -3.5363 |
| Sample 4 | Prog11 inc | | | -2.60948 |
| Sample 5 | Prog4 inc | | | |
| Sample 7 | Prog7 'good' | | | |
| Sample8 | Prog11 'bad' | | | -2.36734 |
| Sample 9 | Prog12 inc | -4.56794 | | -9.65032 |
| Sample 10 | Prog7 'good' | | | -2.58598 |
| Sample 11 | Prog12 'bad' | -3.51655 | | -6.0506 |
| Sample 12 | Prog12 'bad' | -2.14483 | | -3.90297 |
| Sample 13 | Prog11 inc | -2.45251 | | -4.45644 |
| Sample 14 | Prog6 'good' | | | |
| Sample 15 | Prog8 'good' | | | |
| | | | | |

**Table 5.4**. CAM vs. ANM, genes highly correlated with prognosis (0.78) and expression > 2 fold change. Fold change in individual patients and associated patient prognosis score.

In general, more genes showing high correlation with prognosis scores and >1.6 fold changes were identified in the CAM vs. ATM dataset when a 0.79 correlation threshold was applied (Supplementary excel file 5.2). Increasing the stringency, 12 genes were found to have fold changes ≥2 in at least one patient and have the potential to be confirmed using qPCR (Table 5.5). In particular, patient 3 has the highest tumour classification score (13) and the largest number (11/12) of highly correlated genes with +/- fold changes >2.

| | | 1075 | 1535 | 2627 | 3710 | 3915 | 5327 |
|---|---|---|---|---|---|---|---|
| Sample | | CTSC | CYBA | GATA6 | ITPR3 | LAMC1 | PLAT |
| 1 | Prog5 (G) | | | | | | |
| 2 | Prog11 (B) | | 2.241117 | -2.12066 | | | |
| 3 | Prog13 (B) | 2.633758 | 2.137527 | -27.2294 | -2.37362 | -4.57122 | 3.376451 |
| 4 | Prog11 (I) | -2.18877 | | | | | |
| 5 | Prog4 (I) | | | | | | |
| 7 | Prog7 (G) | | | | | | |
| 8 | Prog11 (B) | | | 2.949478 | | | |
| 9 | Prog12 (I) | 2.019918 | | -12.0161 | | | 2.019528 |
| 10 | Prog7 (G) | | | | | | |
| 11 | Prog12 (B) | | | 4.273957 | | | |
| 12 | Prog12 (B) | | | 2.419692 | | | |
| 15 | Prog8 (G) | | | | | | |
| | | 8519 | 51678 | 63898 | 79090 | 150864 | 100128893 |
| Sample | | Ifitm1 | MPP6 | SH2D4A | TRAPPC6A | fam117b | LOC 100128893 |
| 1 | Prog5 (G) | | | | | | |
| 2 | Prog11 (B) | | | -2.60606 | 3.876139 | | -4.60362 |
| 3 | Prog13 (B) | 6.179804 | -2.45994 | -8.46253 | | 2.613945 | -9.88329 |
| 4 | Prog11 (I) | -2.42006 | | | | | |
| 5 | Prog4 (I) | | | | | | |
| 7 | Prog7 (G) | | | | | | |
| 8 | Prog11 (B) | | | | | -2.06323 | |
| 9 | Prog12 (I) | | | -2.07144 | | 2.365152 | -4.03056 |
| 10 | Prog7 (G) | | | | | | |
| 11 | Prog12 (B) | 2.459117 | | 2.05854 | | 2.55737 | 2.35559 |
| 12 | Prog12 (B) | | | | | | |
| 15 | Prog8 (G) | | | | | | |

**Table 5.5.** CAM vs. ATM, genes highly correlated with prognosis (0.79) and expression greater than 2 fold change. Fold change given for individual patients and associated patient prognosis score. Prognosis categories abbreviated as G= 'good', B= 'bad' and I = inconclusive.

The transcription factor GATA6 (2627) is expressed differently in the two 'bad' patient subgroups, it is particularly down regulated in patients 3 and 9 from 'bad' subgroup A, with the worst patient tumour classification score but is up-regulated in 'bad' subgroup B containing patient samples 8, 11 and 12. GATA6 is amplified in pancreatic cancer, contributing to the oncogenic phenotype (Kwei et al., 2008). In colon cancer, GATA6 was also shown to be up-regulated, causing an indirect decrease in apoptosis. Increased expression of GATA6 causes a decrease in the LOX-1 protein, which in turn decreases Fas induced apoptosis (Shureiqi et al., 2007).

Cathepsin C (CYSC) is a cysteine proteinase up-regulated in two 'bad' patients from subgroup A. Cathepsin C is able to activate other cysteine proteinases, together they cleave proteins, degrade the extracellular matrix and increase inflammation, and are known to be up-regulated in many cancers (Mohamed and Sloane, 2006). Interestingly, the expression of plasminogen activator PLAT is increased in 'bad' subgroup A, patients 3 and 9. PLAT is a tissue type plasminogen activator, a secreted serine protease, which converts the pro-enzyme plasminogen to plasmin. Plasmin is a protease which may increase cell migration and remodelling and is known to be up-regulated in a range of tumours, including, glioma (Sandstrom et al., 1999), melanoma (Meissauer et al., 1991), pancreatic (Paciucci et al., 1998) and breast cancer (Chernicky et al., 2005). MPP6, a membrane associated kinase, involved in tumour suppression (Tseng et al., 2001) is decreased in patient sample 3 with the worst prognosis score. Inositol 1,4,5-trisphosphate receptor (ITPR3) is responsible for the release of calcium from intracellular stores, and is up-regulated in two patients from 'bad' subgroup A. SH2D4A is a member of the SH-2 signalling protein family and has been suggested as a possible anti-cancer drug, as it binds directly to the oestrogen receptor, inhibiting oestrogen-induced cell proliferation (Li et al., 2009). In keeping with these findings, SH2D4A has been down regulated in all patients within subgroup A.

Additional patient information became available (Table 5.3) during the course of this study, as a result it was suggested that lymph node metastasis may accurately predicts patient prognosis scores. Therefore, in addition to using patient prognosis scores, correlation analysis was conducted to identify genes whose expression profiles correlate with lymph node metastasis. Firstly, for both datasets, CAM vs.

ANM and ATM vs. ANM, no clear separation was shown to differentiate patients with little or extensive lymph node metastasis, when correlation analysis was conducted comparing lymph node metastasis to gene expression (data not shown). Previously, correlating patient Prognosis scores to gene expression profiles for the CAM vs. ATM myofibroblast dataset revealed two distinct 'bad' patient subgroups. We were therefore interested to determine if the same trend was observed when correlating lymph node metastasis to gene expression. In agreement with previous correlation analysis, patients with worse lymph node metastasis exhibited more extreme changes in gene expression profiles (Figure 5.8A). Each gene was plotted against its correlation score, which ranges between -1 and +1 (Figure 5.8B). A correlation threshold of 0.70 was found to be the optimum positive correlation threshold to apply, resulting in 280 genes (Supplementary excel file 5.3) having larger fold changes in patients with increased lymph node metastasis (Figure 5.8C). Of those positively correlating, 109 had gene expression profiles > 2 fold change in at least one patient (Supplementary excel file 5.4). Genes did not negatively correlate with prognosis (data not shown).

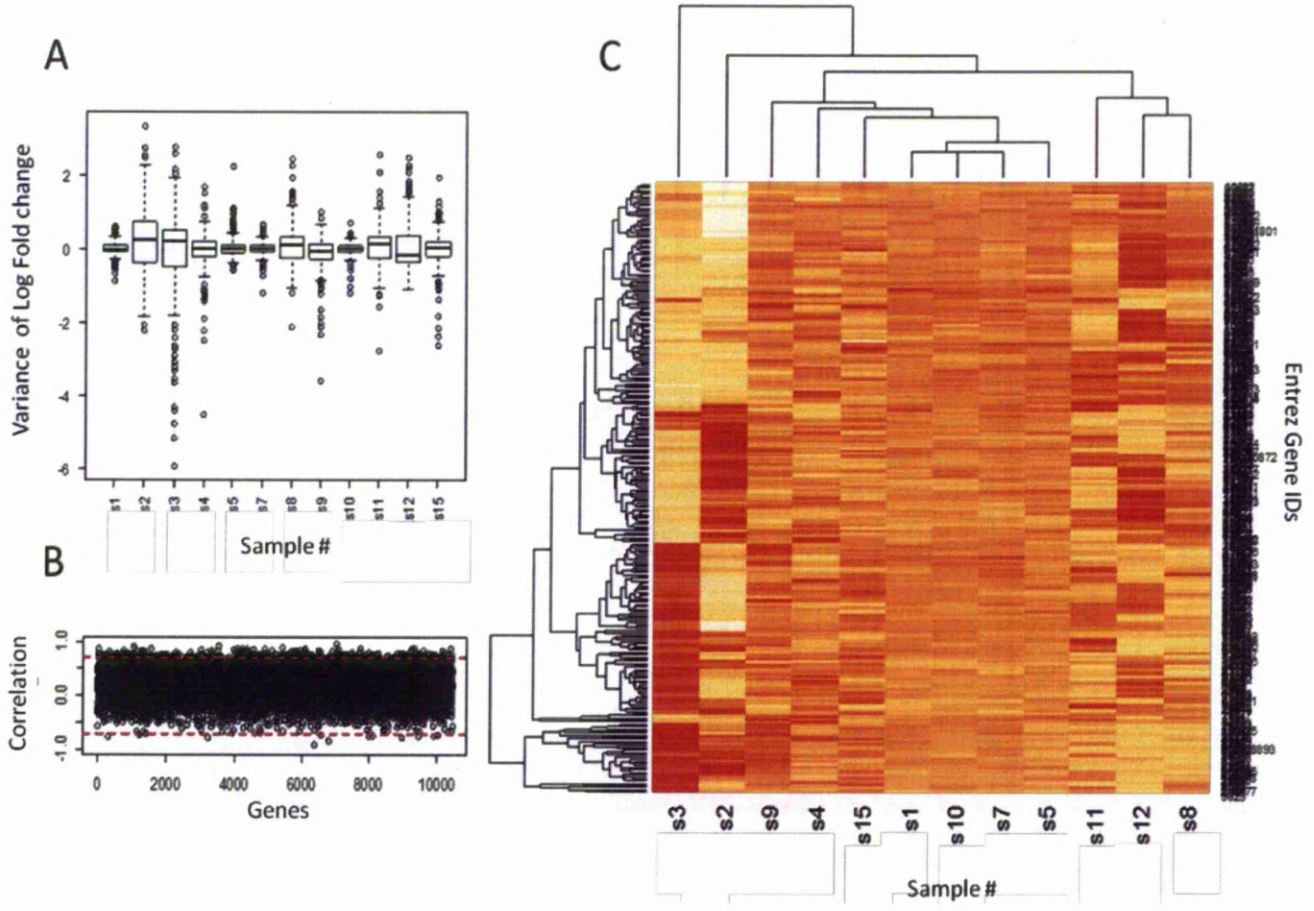



**Figure 5.8.** CAM vs. ATM, lymph node correlation analysis A) Plot representing the variance of gene fold changes in individual patients (patient sample numbers given) B) Spearman's rank correlation threshold ≥0.7 is selected as the optimum value, as depicted by red dashed line. Genes falling above or below red dashed lines are defined as correlated and represented in the heat map in section c. C) Heat map representing the genes positively correlated with expression.

Genes that were found to be positively correlated with lymph node metastasis were then compared to genes that show positive correlation with patient prognosis scores. 28 genes were found to positively correlate with both prognosis score and lymph node metastasis (Table 5.6), of those genes 9 had fold changes ≥2. Therefore, only 3 genes which positively correlated with patient prognosis scores and had fold changes ≥2 were not shown to also correlate with lymph node metastasis.

| Entrez Gene ID | Gene name | Entrez Gene ID | Gene name |
|---|---|---|---|
| 51 | ACOX1 | 23641 | LDOC1 |
| 166 | AES | 51082 | POLR1D |
| 1075 | CTSC | 51647 | FAM96B |
| 1535 | CYBA | 51678 | MPP6 |
| 1635 | DCTD | 51759 | C9orf78 |
| 2627 | GATA6 | 56927 | GPR108 |
| 3915 | LAMC1 | 63898 | SH2D4A |
| 4677 | NARS | 79090 | TRAPPC6A |
| 5327 | PLAT | 79180 | EFHD2 |
| 6223 | RPS19 | 84056 | KATNAL1 |
| 8666 | EIF3G | 114818 | KLHL29 |
| 10581 | IFITM2 | 127281 | C1orf93 |
| 11153 | FICD | 389541 | C7orf59 |
| 23107 | MRPS27 | 100128893 | LOC100128893 |

**Table 5.6**. Genes whose expression is positively correlated with both patient prognosis and lymph node metastasis. Genes highlight in grey, have fold changes ≥2 in at least one patient.

### 5.1.4 Metacore analysis

#### 5.1.4.1 Data processing

After removal of outlying patient samples identified during principal component analysis (Section 5.3.1.1) and re-grouping of 'good' and 'bad' patients, lists of background detected gene lists and significantly changed genes are displayed in Table 5.7.

| Dataset | Background gene list | Significantly changed Genes (p≤0.05) |
|---|---|---|
| CAM vs. ANM | 10,618 | 2,298 |
| ATM vs. ANM | 10,818 | 3,881 |
| CAM vs. ATM | 10,678 | 1,892 |
| 'good' CAM vs. ANM | 10,632 | 313 |
| 'bad' CAM vs. ANM | 10,568 | 1415 |

**Table 5.7** Numbers of background detected gene lists and significantly changed genes (p≤0.05) for each dataset.

The p-values of all significantly changed gene lists (apart from the CAM vs. ATM dataset) were corrected using the Benjamini-Hochberg false discovery rate (FDR) algorithm. In the CAM vs. ATM dataset no genes passed a FDR ≤0.05 and the gene list passing, an un-corrected p-value ≤0.05 was much smaller than the other two datasets. This is most likely due to the high degree of similarity in gene expression profiles expected for samples taken from the same tissue of the same individual, as highlighted by principal component analysis (Figure 5.2). Of the 1,892 genes found to be significantly changed in the CAM vs. ATM dataset, 132 have a fold change greater than 1.6, and 49 have a fold change greater than 2, Metacore™ analysis below will be used to select the optimum threshold to apply. The number of genes classed as statistically changed in the 'bad' patient sub-set is far higher than that observed in the 'good' patient sub-set, suggesting that samples derived from 'bad' sub-group patients are consistently more altered when compared to normal samples than those derived from patients with 'good' prognosis scores. Metacore™ GeneGo analysis was performed on five different datasets; CAM vs. ANM; ATM vs. ANM; CAM vs. ATM and 'good' or 'bad' CAM vs. ANM datasets. Twenty eight pathways were found to be over-represented (p≤0.05) in the CAM vs. ANM dataset

in comparison to 35 pathways that were over-represented ($p \leq 0.05$) in the ATM vs. ANM dataset

Within the CAM vs. ANM dataset, a large proportion of over-represented pathways (10/28), are involved in regulation of the cell-cycle, of which all differentially regulated genes are down-regulated, suggesting a decrease in cell proliferation. Cell cycle dependent kinases, genes required for spindle-chromatin attachment, the APC complex, which induces exit from mitosis and are all down-regulated throughout, suggesting a decrease in progression through the cell cycle stages and an overall decrease in proliferation. Genes involved in DNA damage induced apoptosis and DNA damage checkpoint arrest are down-regulated, leading to loss of internal safety controls upon potentially dangerous DNA damage. Down-regulated DNA damage caretaker proteins, include Brca1 and Brca2. Brca1 and Brca2 were both linked to breast cancer in 1994 (Hall et al., 1990; Wooster et al., 1994) and since then have been shown to be important caretaker genes, repairing DNA, remodelling chromosomes and orchestrating DNA damage signalling (Duncan et al., 1998; Durant and Nickoloff, 2005). The over-representation of ketone body synthesis pathway is associated with increased expression of key enzymes regulating the conversion of acetyl-CoA into the soluble ketone body, acetoacetate (Williamson et al., 1971).

Within the ATM vs. ANM dataset, interestingly the top pathway hit was cholesterol biosynthesis, with all of the differentially regulated genes up-regulated. As for ketone body synthesis (over-represented in the CAM vs. ANM dataset), cholesterol biosynthesis can also be synthesised from acetyl Co-A. This apparent difference in

the metabolism of acetyl-coA within CAMs and ATMs is interesting and will be investigated further. A smaller number of cell cycle pathways (6/35) are over-represented within the ATM vs. ANM dataset, compared to the CAM vs. ANM dataset. Although the trend is the same, with all differentially regulated genes hitting the pathway down regulated, suggesting an over-all decrease in cell proliferation. Another similar trend to the CAM vs. ANM dataset, is the large number of genes that are down-regulated in response to DNA damage, usually involved in cell cycle arrest, such as Brca1. Unique pathways coming through within the ATM vs. ANM dataset include an increase in the expression of genes controlling smooth muscle contraction, which seems appropriate for the myofibroblast cell type and an overall increase in the expression of genes involved in a range of amino acid metabolism pathways.

To establish an appropriate fold change cut-off to define a statistically significant CAM vs. ATM gene list, Metacore™ analysis was carried out using both a 1.6 and 2 fold change cut-off. Table 5.8 displays statistically over-represented ($p \leq 0.05$) pathways, after applying a 0, 1.6 or 2-fold change cut-off. As expected, the more stringent the fold change cut-off applied the smaller the resulting gene list and fewer pathways were found to be statistically over-represented. Nearly half of the pathways statistically over-represented ($p \leq 0.05$) in the 2 fold change gene list were also over-represented ($p \leq 0.05$) in the 1.6 fold change gene list, indicating that completing a 1.6 fold change cut-off wouldn't impair pathway enrichment analysis.

| No fold change cut-off | 1.6 fold change cut-off | 2 fold change cut-off |
|---|---|---|
| Apoptosis and survival_Anti-apoptotic TNFs/NF-kB/Bcl-2 | Bile Acid Biosynthesis | Bile Acid Biosynthesis |
| Apoptosis and survival_Endoplasmic reticulum stress response | Bile Acid Biosynthesis / Rv | Bile Acid Biosynthesis / Rodent version |
| Apoptosis and survival _Inhibition of ROS-induced apoptosis by 17beta-estradiol | Androstenedione and testosterone biosynthesis and metabolism p.2/ Rv | Development_WNT5A signaling |
| Blood coagulation_GPVI-dependent platelet activation | Androstenedione and testosterone biosynthesis and metabolism p.2 | Development_TGF-beta-dependent induction of EMT via MAPK |
| Butanoate metabolism | Development_TGF-beta-dependent induction of EMT via MAPK | Development_MicroRNA-dependent inhibition of EMT |
| Cardiac Hypertrophy_Ca(2+)-dependent NF-AT signalling in Cardiac Hypertrophy | Development_Regulation of epithelial-to-mesenchymal transition (EMT) | Tyrosine metabolism p.1 (dopamine) |
| Catecholamine metabolism / Human version | Oxidative stress_Angiotensin II-induced production of ROS | Histamine metabolism |
| CFTR folding and maturation (norm and CF) | Development_TGF-beta-dependent induction of EMT via SMADs | |
| Cholesterol Biosynthesis | Blood coagulation_Blood coagulation | |
| Cytoskeleton remodeling_Reverse signaling by ephrin B | Cell adhesion_Endothelial cell contacts by junctional mechanisms | |
| Development_FGF-family signaling | Aspartate and asparagine metabolism | |
| Development_PEDF signaling | Cell adhesion_Integrin inside-out signaling | |
| Development_Role of IL-8 in angiogenesis | Immune response_IL-4 - antiapoptotic action | |
| Development_TGF-beta-dependent induction of EMT via MAPK | | |
| Development_TGF-beta-dependent induction of EMT via RhoA, PI3K and ILK. | | |
| Development_TGF-beta-dependent induction of EMT via SMADs | | |
| Immune response_NFAT in immune response | | |
| Neurophysiological process_ACM regulation of nerve impulse | | |
| NGF activation of NF-kB | | |
| Nitrogen metabolism | | |
| Nitrogen metabolism/ Rodent version | | |
| Propionate metabolism p.2 | | |
| Vitamin B7 (biotin) metabolism | | |
| wtCFTR and delta508-CFTR traffic / Generic schema | | |

**Table 5.8.** Comparison of statistically over-represented ($p \leq 0.05$) Metacore pathways, upon the application of genes passing no fold change cut-off, a 1.6 fold change cut-off and a 2-fold change cut-off. The red highlighted pathway in common in all three thresholds, grey pathways are common in the 1.6 and 2-fold change pathway, whilst the taupe pathway is common in the 1.6 and no fold change pathway.

Upon closer assessment, due to the smaller differentially regulated 2 fold cut-off gene list, very few genes are mapped to the over-represented pathways, with only 1 gene being mapped in some cases. This makes it difficult to assess pathway assignment accurately; therefore the 2 fold over-represented pathways alone was deemed uninformative. It may be sensible to complete pathway analysis with both gene lists, using the over-represented pathways from the 2 fold cut-off as the high confidence set and the over-represented pathways from the 1.6 fold cut-off list to provide larger numbers of differentially regulated genes mapped to pathways. The red highlighted pathway, the development TGF-beta-dependent induction of EMT via MAPK, is common in all three-fold change cut-off pathway lists. Transforming growth factor itself is increased but only one other member of the pathway is differentially regulated, therefore its role is unclear. Although TGF-β is known to stimulate differentiation of a fibroblast into the activated myofibroblast phenotype therefore it is interesting that this pathway has appeared statistically enriched in CAMs compared to the ATMs. In addition, EMT takes place within the cancer cells, therefore this pathway would not be expected to take place in the myofibroblast themselves. The fact that TGF-β expression is increased in CAMs compared to ATMs suggests that CAMs may stimulate EMT though increased secretion of TGF-β.

Understanding the unique pathways over-represented in the 'good' and 'bad' prognosis sub-groups may help us to understand the biological processes that contribute to patient prognosis. Whilst understanding the common pathways over-represented in the 'good' and 'bad' prognosis groups may help us to understand the biological processes that are inherently changed throughout the different stages of tumour development. Firstly, 30 pathways are over-represented in the 'bad'

prognosis score, CAM vs. ANM dataset, which is far more than the 12 pathways over-represented in the 'good' prognosis score, CAM vs. ANM dataset. This increase in the number of statistically significantly over-represented pathways could be due to the larger number of significant changed genes found in the 'bad' prognosis group. Twenty six pathways are uniquely over-represented in the 'bad' dataset (Table 5.9), there is a general down regulation in all genes involved in the four over-represented cell-cycle pathways, suggesting an overall decrease in cell proliferation, consistent with the overall datasets. Interestingly, three separate over-represented pathways are involved in DNA damage detection and down-regulation of caretaker genes, such as Brca1. No such signature arises within the 'good' prognosis dataset, and may suggest a genetic difference, which causes a more aggressive tumour. As also shown in the CAM vs. ANM dataset, mitochondrial ketone body biosynthesis is over-represented. With enzymes converting acetyl-CoA into the soluble ketone, acetoacetate, displaying increased expression.

Twelve pathways are uniquely over-represented in the 'good' dataset (Table 5.10). With the top four unique pathways are involved in the cell cycle, Cell cycle_Start of DNA replication in early S phase passes a FDR corrected p-value of 0.05, displaying four genes with reduced expression, and in fact, genes within all the unique cell cycle pathways are down regulated. In addition, the genes differentially regulated throughout the majority of the unique 'good' Metacore™ pathways seem to be either CDKs or the transcription factor E2F1. However the 'Translation_(L)-selenoaminoacids incorporation in proteins during translation' pathway identifies the increased expression of a range of anti-oxidant enzymes, whose increase in expression is often in response to oxidative stress. This increase in expression of

protective anti-oxidant enzymes within patients with 'good' prognosis scores could be a protective mechanism, decreasing the aggressiveness of the tumour. Within the G-protein signalling K-RAS regulation pathway, K-RAS is down regulated, suggesting a decreased propagation signal from growth factors within 'good' patient subgroups.

| |
|---|
| Cell cycle_Start of DNA replication in early S phase |
| Cytoskeleton remodeling_Role of PKA in cytoskeleton reorganisation |
| Cell cycle_Transition and termination of DNA replication |
| DNA damage_ATM/ATR regulation of G1/S checkpoint |
| Fatty Acid Omega Oxidation |
| Nicotine signaling in dopaminergic neurons, Pt. 2 - axon terminal |
| Transport_ACM3 in salivary glands |
| Leucine, isoleucine and valine metabolism/ Rodent version |
| Leucine, isoleucine and valine metabolism.p.2 |
| DNA damage_DNA-damage-induced responses |
| Serotonin modulation of dopamine release in nicotine addiction |
| Cardiac Hypertrophy_Ca(2+)-dependent NF-AT signaling in Cardiac Hypertrophy |
| Cell adhesion_Cadherin-mediated cell adhesion |
| Vitamin E (alfa-tocopherol) metabolism |
| DNA damage_Role of Brca1 and Brca2 in DNA repair |
| Neurophysiological process_Glutamate regulation of Dopamine D1A receptor signaling |
| Cytoskeleton remodeling_ACM3 and ACM4 in keratinocyte migration |
| Mitochondrial ketone bodies biosynthesis and metabolism |
| Apoptosis and survival_DNA-damage-induced apoptosis |
| Apoptosis and survival_Granzyme A signaling |
| Cell cycle_Sister chromatid cohesion |
| Immune response_Antigen presentation by MHC class I |
| G-protein signaling_Regulation of CAMP levels by ACM |
| Leukotriene 4 biosynthesis and metabolism |
| Cytoskeleton remodeling_Alpha-1A adrenergic receptor-dependent inhibition of PI3K |
| Cholesterol and Sphingolipids transport / Recycling to plasma membrane in lung (normal and CF) |

**Table 5.9.** Uniquely over-represented Metacore pathways in the 'bad' tumour classification set.

| |
|---|
| Cell cycle_Cell cycle (generic schema) |
| DNA damage_Inhibition of telomerase activity and cellular senescence |
| Cell cycle_Regulation of G1/S transition (part 2) |
| Cell cycle_ESR1 regulation of G1/S transition |
| Translation_(L)-selenoaminoacids incorporation in proteins during translation |
| G-protein signaling_K-RAS regulation pathway |
| Cell cycle_Influence of Ras and Rho proteins on G1/S Transition |
| Translation _Regulation of translation initiation |
| Polyamine metabolism |
| HETE and HPETE biosynthesis and metabolism |
| Signal transduction_JNK pathway |
| Signal transduction_ERK1/2 signaling pathway |

**Table 5.10** Uniquely over-represented Metacore pathways in the 'good' tumour classification set.

Four pathways are commonly over-represented within both the 'bad' and 'good' dataset. Differentially regulated genes involved in apoptosis and survival_Granzyme A signalling, are down regulated, suggesting a decrease in apoptosis across patients with varying degrees of tumour progression.

| | 'bad' p-value | 'good' p-value |
|---|---|---|
| Gamma-secretase proteolytic targets | 0.000524 | 1.612E-07 |
| Gamma-Secretase regulation of neuronal cell development and function | 0.000259 | 0.00009861 |
| Proteolysis_Putative ubiquitin pathway | 0.03357 | 0.03753 |
| Regulation of lipid metabolism_Stimulation of Arachidonic acid production by ACM receptors | 0.01317 | 0.01808 |

**Table 5.11.** Common over-represented Metacore pathways within the 'good' and 'bad' tumour classification set. Red p-values represent the most significant over-representation result between the 'good' and 'bad' prognosis groups.

### 5.1.4.2 Transcription factor analysis

The Metacore™ one click analysis function was used to identify transcription factors that were statistically changed (p-values≤0.05) within the CAM vs. ANM, ATM vs. ANM and CAM vs. ATM datasets. Differentially expressed (p≤0.05) transcription factors identified in all three datasets are shown in Table 5.12. In total 51 transcription factors were identified in the CAM vs. ANM dataset, 35 in the CAM vs. ATM and 95 in the ATM vs. ANM patients.

Numbers of differentially regulated transcription factors (p≤0.05) common between: CAM vs. ANM; CAM vs. ATM and ATM vs. ANM, are shown in Figure 5.9. Lists of genes represented in each segment are given in Supplementary excel file 5.5

| CAM vs. ANM | CAM vs. ATM | ATM vs. ANM | |
|---|---|---|---|
| ETV3 | STAT1 | ETV3 | TFE3 |
| ETS1 | GTF2I | XBP1 | TWIST1 |
| YY1 | EGR1 | EPAS1 | FUBP1 |
| RXRA | RELA | EGR1 | CTCFL |
| HCFC1 | SOX9 | YY1 | SMAD3 |
| ATF3 | BCL6 | CBFB | MAFB |
| E2F1 | ETV5 | RXRA | ZNF350 |
| FUBP1 | PURA | HCFC1 | KLF2 |
| HES1 | MZF1 | CTCF | ZNF83 |
| ZNF263 | HHEX | ATF3 | ETV1 |
| ZNF217 | SMAD5 | E2F1 | MYEF2 |
| IRF9 | SMAD3 | RFX5 | PPARA |
| CREB1 | FOXD1 | ARNT2 | SP1 |
| FOSL1 | ILF3 | BCL6 | ETS1 |
| CLOCK | MEF2C | IRF2 | FOXO3 |
| MTF1 | RUNX1 | ZNF148 | TEAD1 |
| GLI2 | DDIT3 | ETV5 | NFAT5 |
| NFATC3 | CREB3 | HES1 | SMAD5 |
| TFAM | C2orf3 | SATB1 | KLF13 |
| PA2G4 | KLF6 | ZNF263 | GTF2A1 |
| HMGB2 | PBX1 | FOXJ2 | CEBPG |
| KLF6 | MEF2A | ZNF217 | CREB1 |
| TCF3 | KLF3 | NR1H3 | STAT2 |
| MEF2C | ATF1 | GABPB1 | TFDP2 |
| MAX | GMEB2 | NR2C1 | AR |
| DDIT3 | ZRANB2 | TCF21 | FOXK2 |
| DBP | ETS1 | CLOCK | ZBTB33 |
| TCF4 | NFAT5 | PRDM2 | ATF6 |
| NFE2L1 | NFATC3 | POU6F1 | MTF1 |

| | | | |
|---|---|---|---|
| TEAD1 | FOSL2 | CNBP | E2F7 |
| CTCFL | GLI2 | GLI2 | TWIST2 |
| ILF3 | MXD1 | RBPJ | SIX4 |
| E4F1 | RUNX2 | PA2G4 | NFIA |
| KLF2 | SREBF2 | HMGB2 | ARNT |
| FOXO3 | TAF9 | KLF6 | GMEB1 |
| NFAT5 | | NR2F2 | NFATC4 |
| STAT2 | | TCF3 | ELK4 |
| TFDP2 | | MEF2C | ZFHX3 |
| FOXK2 | | KLF5 | GTF2A2 |
| ZBTB33 | | MAX | ZNF224 |
| ATF6 | | MAF | THRA |
| LRRFIP1 | | DDIT3 | NPAS2 |
| EGR1 | | TEAD3 | TEAD4 |
| E2F7 | | NR2F1 | SREBF2 |
| SIX4 | | DBP | NFATC1 |
| ETV5 | | MZF1 | MAZ |
| SMAD5 | | NFATC3 | TCF4 |
| ELK4 | | GTF2I | |
| THRA | | | |
| MZF1 | | | |
| SREBF2 | | | |

**Table 5.72.** Differentially expressed transcription factors identified in, CAM vs. ANM (FDRp≤0.05), CAM vs. ATM (p≤0.05) and ATM vs. ANM (FDR p≤0.05) datasets. No fold change cut-off was applied in this analysis.

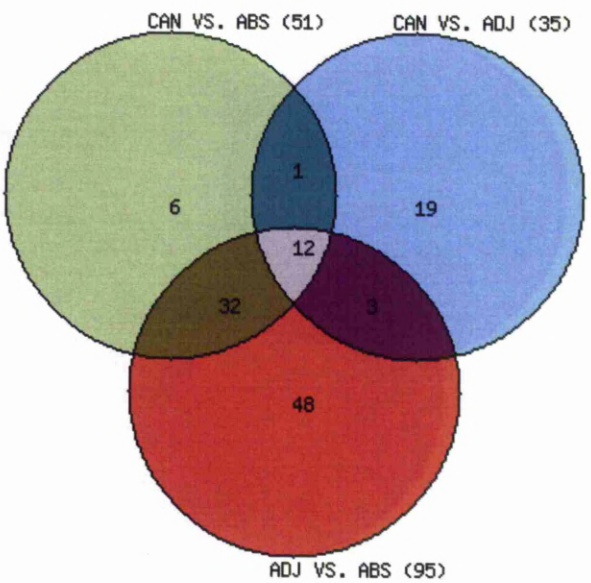



**Figure 5.9.** Venn diagram displaying the over-lap of differentially regulated transcription factors (p≤0.05) in the CAM vs. ANM, CAM vs. ATM and ATM vs. ANM datasets.

Gene expression profiles of the differentially regulated transcription factors were visualised in Partek® using hierarchal clustering analysis. The CAM vs. ANM dataset (Figure 5.10) shows the gene expression profiles of differentially regulated transcription factors at $p \leq 0.05$ (Figure 5.10A) or $p \leq 0.05$ and 1.6 fold change (Figure 5.10B), which separates the cancer samples (red) and absolute normal samples (green) perfectly. Transcription factors are also separated into two clear groups, 5 transcription factors are down regulated in the cancer-derived samples compared to the normal and 9 transcription factors are up regulated in the cancer samples compared to the normal. These two facts together, result in a clear four quadrant hierarchical clustering plot.

Comparable analysis of the ATM vs. ANM dataset (Figure 5.11) revealed very similar findings to the CAM vs. ANM dataset, in that, the gene expression of differentially regulated transcription factors separate ATM and ANM samples perfectly and also segregate transcription factors into two groups; those that are up regulated in ATMs and those that are down regulated in ATMs. When no-fold change cut-off is applied, the numbers of transcription factors that are over and under regulated in adjacent samples is approximately equal. Raising the fold change cut-off to 1.6 results in similar findings as seen in the CAM vs. ANM dataset, with the majority of transcription factors being down regulated in adjacent samples. For the CAM vs. ATM dataset (Figure 5.12), hierarchical clustering analysis of differentially regulated transcription factors, does not reveal clear separation of cancer and adjacent samples. A 1.4 fold change resulted in a slightly better, but not perfect, separation of samples, highlighted by the lack of clear red and blue quadrants. Transcription factors with the highest fold changes, between 1.4 -1.55 fold-change include;

DDIT3, SOX9, RUNX1, KLF6 and SMAD3. Overall, upon application of a range of fold change cut-offs it has become apparent that transcription factor gene expression fold changes are not large across any of the datasets.

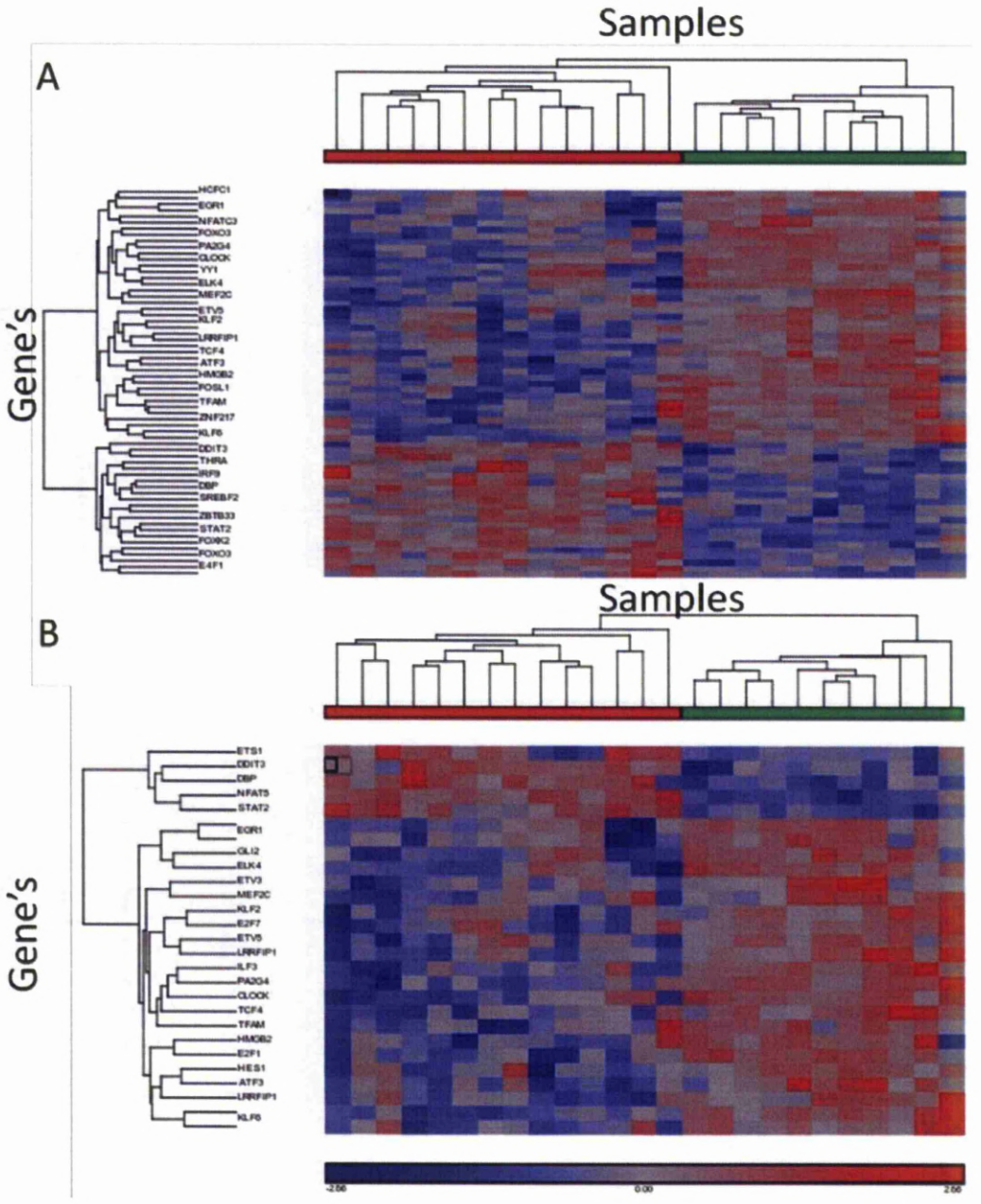



**Figure 5.10**. Transcription factors identified in the CAM vs. ANM differentially regulated gene lists. (A) Statistically significant transcription factors FDR p≤0.05. (B) Statistically significant transcription factors FDR p≤0.05 and >1.6 fold change. Individual patients on the x-axis and transcription factors on the y-axis. Red patients represent cancer samples, whilst green patients represent absolute normal samples.

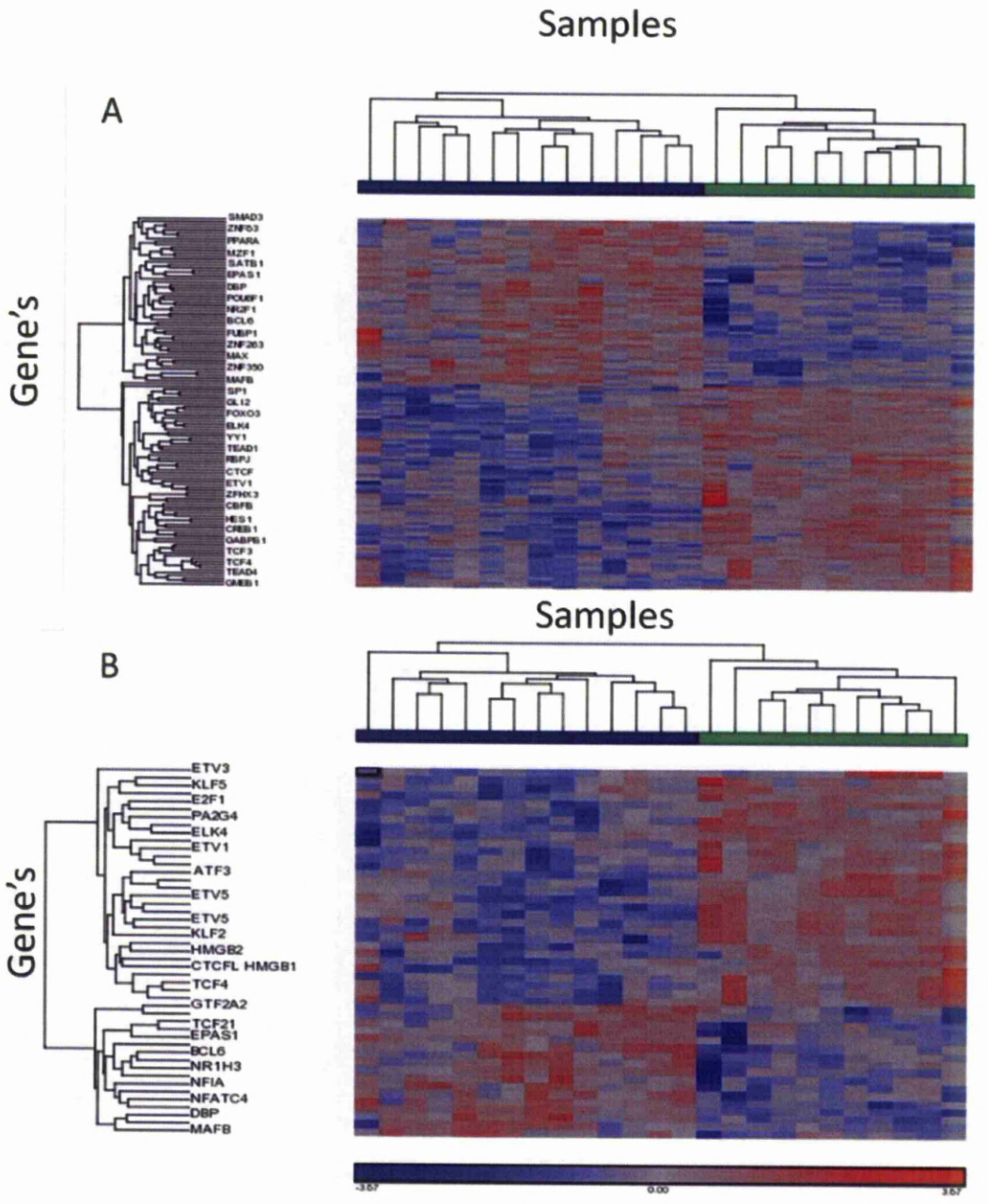



**Figure 5.11.** Transcription factors identified in the ATM vs. ANM differentially regulated gene lists. (A) Statistically significant transcription factors FDR p≤0.05. (B) Statistically significant transcription factors FDR p≤0.05 and >1.6 fold change.

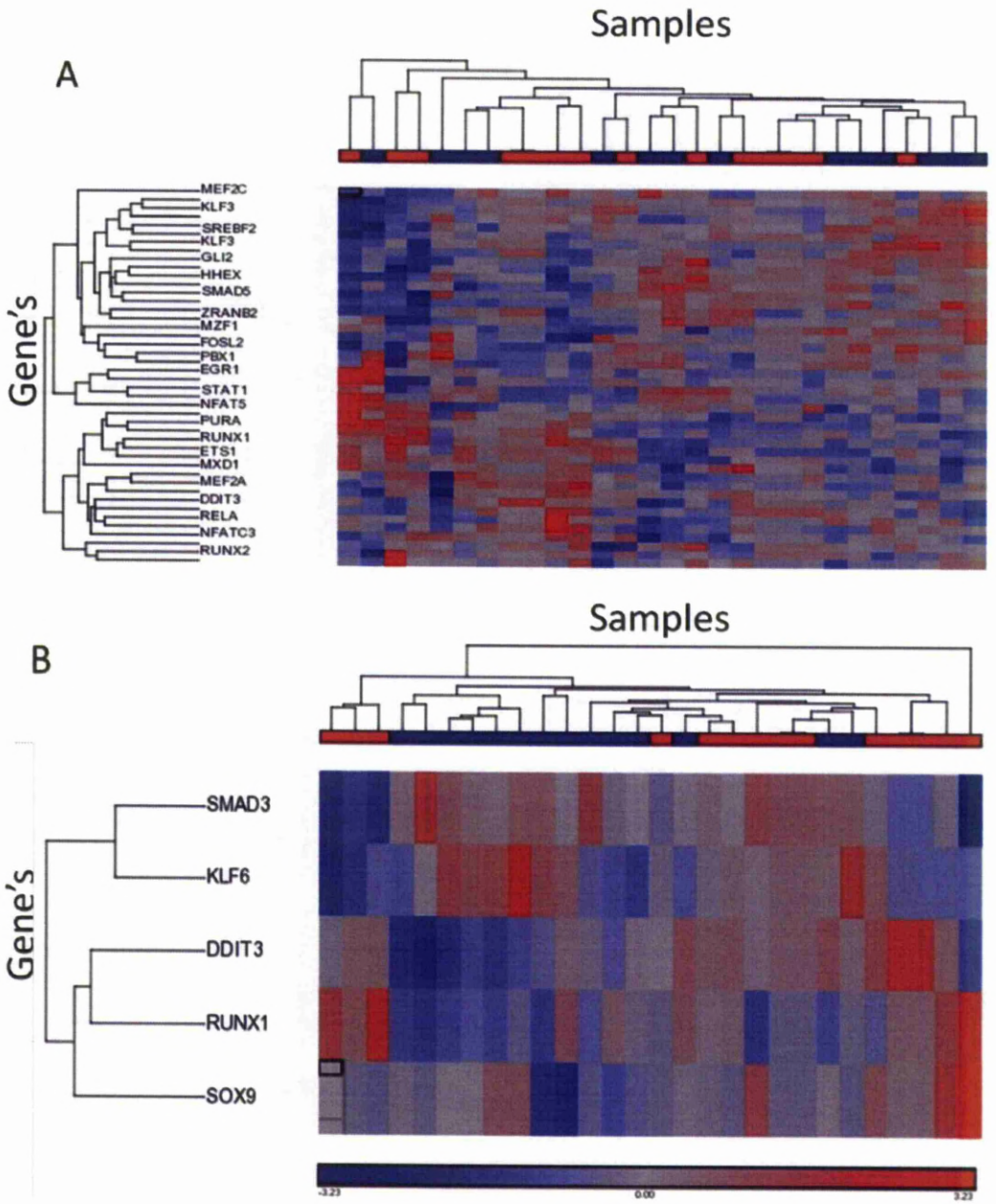
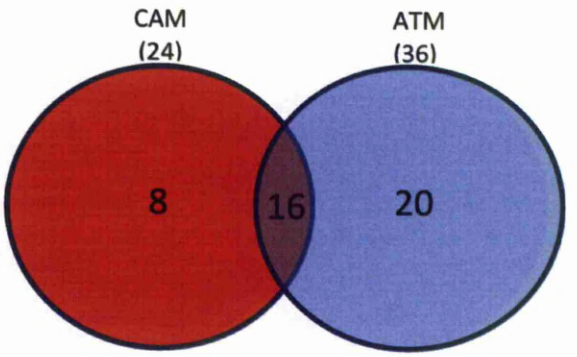



**Figure 5.12.** Transcription factors identified in the CAM vs. ATM differentially regulated gene lists. (A) Statistically significant transcription factors p≤0.05. (B) Statistically significant transcription factors p≤0.05 and >1.4 fold change.

As the differentially regulated transcription factors hierarchal clustering analysis failed to separate CAM vs. ATM samples clearly, it seemed sensible to compare transcription factors in just the CAM vs. ANM and ATM vs. ANM datasets, as the

gene expression of these transcription factors gave clear separation of sample types. Figure 5.13 shows the similarity of transcription factors with fold changes >1.6 that are differentially regulated in CAM and ATM samples. Of those 16 transcription factors, 13 have larger negative fold changes in the ATM than the CAM.




| Cancer | Common | Adjacent |
|---|---|---|
| *DDIT3* | *DBP* | *TCF21* |
| *ETS1* | *STAT2* | *MAF* |
| *NFAT5* | *PA2G4* | *NFIA* |
| *TFAM* | *ELK4* | *MAFB* |
| *ILF3* | *E2F1* | *EPAS1* |
| *GLI2* | *ETV3* | *ZNF224* |
| *LRRFIP1* | *CLOCK* | *TWIST1* |
| *MEF2C* | *TCF4* | *NFATC4* |
| | *HMGB2* | *NR1H3* |
| | *KLF2* | *NR2F1* |
| | *KLF6* | *GTF2A2* |
| | *ETV5* | *MAX* |
| | *EGR1* | *SATB1* |
| | *E2F7* | *BCL6* |
| | *ATF3* | *CTCFL* |
| | *HES1* | *MAZ* |
| | | *ZFHX3* |
| | | *MYEF2* |
| | | *KLF5* |
| | | *ETV1* |

**Figure 5.13.** The intersection of differentially regulated ($p \leq 0.05$ and 1.6 fold change) transcription factors in cancer and adjacent samples.

#### 5.1.4.2.1 Transcription factors expressed in 'good' and 'bad' patient subsets

Transcription factors that were significantly changed (p≤0.05) within the 'good' and 'bad' patient prognosis sub-sets, CAM vs. ANM datasets. Transcription factors were identified using the Metacore 'one click analysis' transcription factor tool. Thirty-three transcription factors were changed in the 'bad' prognosis dataset, far more than the 7 changed in the 'good' patient prognosis set. Transcription factors differentially regulated in the 'good' patient prognosis group included *RBL1*, *E2F1*, *CLOCK*, *MTF1*, *PA2G4*, *STAT2* and *SMAD5*, all of which, apart from *STAT2* were down regulated. Transcription factors differentially regulated in the 'bad' patient prognosis group are listed in Table 5.13

| 'Bad' patient prognosis transcription factors | |
|---|---|
| *RBL1* | *MTF1* |
| *HINT1* | *BPTF* |
| *SCP2* | *PDIA3* |
| *EGR1* | *PA2G4* |
| *RXRA* | *MEF2C* |
| *E2F1* | *MAX* |
| *FUBP1* | *TCF4* |
| *ETV5* | *CTCFL* |
| *FOXJ2* | *KLF6* |
| *NF1* | *FOXO3* |
| *FOSL1* | *NFAT5* |
| *BRCA1* | *STAT2* |
| *NFE2L3* | *UHRF1* |
| *CLOCK* | *TFDP2* |
| *E2F7* | *ATF6* |
| *ARNT* | *NACC1* |
| *NFATC4* | |

**Table 5.13.** Differentially regulated (p≤0.05) transcription factors in the cancer vs. absolute 'bad' patient prognosis group.

Comparison of differentially regulated transcription factors in the 'good' and 'bad' patient prognosis groups would be interesting to identify transcription factors that could potentially regulate the range of significantly regulated genes associated with particular prognosis scores. A direct comparison of transcription factors identified in each group is shown in Figure 5.14.

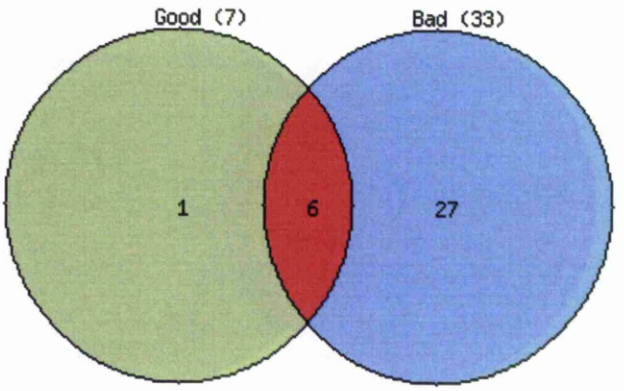



**Figure 5.14.** Intersection of differentially regulated (p≤0.05) transcription factors in the CAM vs. AN 'good' and 'bad' patient prognosis groups. Numbers outside of the circles represent total numbers of differentially regulated genes.

Of the six common differentially expressed transcription factors in both the 'good' and 'bad' prognosis sub-groups, all but *STAT2* were down regulated and 4/6 had a more extreme log fold changes in the 'bad' patient prognosis set. The transcription factors displaying this progressive change include, *E2F1, PA2G4, RBL1* and *STAT2*. *E2F1* promotes the expression of cell-cycle and tumour suppressor genes. Expression of pro-apoptotic genes, are regulated by *E2F1* directly or through *E2F1* transcriptional regulation of *P53* (Polager and Ginsberg, 2009). Progressive down-regulation of *E2F1* within 'bad' patient subgroups suggests a lack of cell-death signals within cancer-associated myofibroblasts, thus able to support tumour

growth in un-desirable environments. *PA2G4* dependent on the isoform expressed is either an oncogene or a tumour repressor (Liu et al., 2006). The longer, nuclear form inhibits apoptosis whilst the shorter cytoplasmic form inhibits cell proliferation. In view of this, progressive down regulation in the 'bad' patient group suggests expression of the shorter cytoplasmic isoform.

*RBL1* binds to the E2F transcription factor family and inhibits their ability to activate cell cycle genes, thus is thought to have roles in tumour suppression by preventing progression through the cell cycle (Munger and Howley, 2002). In addition, *RBL1* acts epigenetically as a transcriptional repressor by recruiting chromatin-modification enzymes to the promoter (Lai et al., 1999). *STAT2* signal transducer and activator of transcription, upon activation *STAT2* enters the nucleus and binds with IFN regulatory factor family protein *p48* (*ISGF3G*), activating transcription (Martinez-Moczygemba et al., 1997). Overall decreased expression of these genes involved in transcriptional regulation, across patients with worsening prognosis scores is in keeping with previous findings, with relation to their roles in supporting tumour progression

### 5.1.5 Epigenetics

Epigenetics is known to play a key role in the differential expression of genes during carcinogenesis, yet the role of epigenetic regulation in cancer-associated myofibroblasts remains poorly characterised. The fact that cancer-associated myofibroblasts seem to retain their ability to confer cancer cell proliferation and migration when isolated as purified primary cultures suggests that they have in some way been re-programmed. Recent evidence suggests that cancer associated

myofibroblasts do not accumulate significant chromosomal abnormalities or mutations as observed for most tumour cells. This raises the possibility that induced changes may be conferred by epigenetics. Therefore, we were interested to see to what extent expression of proteins involved epigenetic regulation may be changed in cancer derived myofibroblasts gene expression, potentially leading to oncogenesis. The prefix-epi stands for on top or in addition to, therefore epigenetics relates to anything other than information provided by the genetic code (Sandoval and Esteller, 2012) . Genes involved in epigenetic regulation can be split into two types, epigenetic chromatin modification enzymes and epigenetic chromatin remodelling factors. Epigenetic chromatin modification enzymes modify the DNA and histones by methylation, de-methylation, phosphorylation, acetlyation, deacetylation and ubiquitination. Although the epigenetic code is complex and combinatorial, epigenetic chromatin-remodelling factors recognise the modified histones or DNA and remodel chromatin to allow accessibility to DNA gene expression. We compiled a list of 164 epigenetic genes, consisting of 83 epigenetic chromatin modification enzymes and 81 epigenetic chromatin-remodelling factors (Supplementary excel file 5.6 as stated on the Quagen PCA arrays and papers stated in methods chapter 2, section 2.9. The 'good' and 'bad' patient prognosis, CAM vs. ANM datasets were analysed for differentially regulated epigenetic genes (Table 5.14). Shared differentially regulated epigenetic genes within the 'good' and 'bad' patient prognosis groups include *PBRM1* and *CHD1*, which both show decreased expression. *PBRM1* acts as a negative regulator of cell proliferation, with knockdowns shown to lead to renal cancer (Varela et al., 2011); therefore, its reduced expression across patients with varying degrees of tumour severity, and

increased reduced expression in 'bad' patient subgroups suggests it as a possible general epigenetic therapeutic target. The chromatin modification organiser *CHD1* alters the structure of the chromatin on DNA, to control the accessibility of the DNA by transcriptional factors. *CHD1* generally works to maintain an open chromatin structure.

| Entrez Gene ID | Gene name | 'Bad' CAM vs. ANM log FC | 'Good' CAM vs. ANM log FC |
|---|---|---|---|
| 23468 | *CBX5* | -0.955037952 | |
| 80854 | *setd7* | -0.864601896 | |
| 2186 | *BPTF* | -0.796579836 | |
| 54556 | *Ing3* | -0.758565437 | |
| 23476 | *brd4* | -0.745729218 | |
| 1108 | *CHD4* | -0.698922229 | |
| 55193 | *pbrm1* | -0.691598035 | -0.566572281 |
| 6597 | *SMARCA4* | -0.69070869 | |
| 84444 | *Dot1l* | -0.679232763 | |
| 1105 | *CHD1* | -0.506866988 | -0.333428964 |
| 5253 | *Phf2* | -0.245098803 | |
| 3275 | *PRMT2* | 0.358612005 | |
| 4152 | *mbd1* | 0.628257285 | |
| 23492 | *CBX7* | 0.709790185 | |
| 10783 | *Nek6* | 0.864250001 | |
| 1488 | *CTBP2* | | 0.435127622 |
| 8726 | *EED* | | -0.447436142 |

**Table 5.14.** Differentially regulated epigenetic genes ($p \leq 0.05$), identified within the 'bad' and 'good' patient prognosis, CAM vs. ANM datasets. Green shading represents down-regulated genes and red shading represents up-regulated genes.

A larger number of epigenetic genes were differentially regulated within the 'bad' patient subgroups, indicating that epigenetic regulation may be more pronounced

in more aggressive or later stages of cancer progression. Histone methytransferase by *SETD7*, is an important pre-initiation step, which epigeneticaly activates insulin and collegenase gene expression (Kwon et al., 2003; Martens et al., 2003), collagenases breakdown collagen within the extracellular matrix. *MBD1* binds to methylated DNA, inhibiting gene transcription, its down-regulation in 'bad' patients, suggests a release of negative control on a subset of genes (Ng et al., 2000). *ING3* is a member of the inhibitor of growth protein family, the tumour suppressor acts through binding to methylated histones, inducing cellular senescence. Reduced expression of *ING5* has been demonstrated in many cancers (Ludwig et al., 2011). Here *ING5* is uniquely reduced within the 'bad' patient prognosis group, and reduced mRNA expression of an alternative *ING* family member has previously been shown to directly relate to poor prognosis in head and neck cancers (Gunduz et al., 2008; Ludwig et al., 2011). *SMARCA4* has an ATP dependent nucleosome-remodelling complex, and many inactivating mutations within this have been shown in cancers (Medina et al., 2008; Sudarsanam and Winston, 2000). Therefore, *SMARCA4* is often referred to as a tumour repressor; acting to repress *ZEB1*, which itself increases E-cadherin expression and initiates epithelial mesenchymal transition (EMT) (Sanchez-Tillo et al., 2010).

As numbers of differentially regulated genes given epigenetic roles are relatively low, 15/1415 within the 'bad' patient subgroup and 4/313 within the 'good' patient subgroup, a 164 gene epigenetic network was generated within the human protein-protein interactome. The mapping of such an epigenetic network on to the human interactome provides us with essential protein-protein interaction information of

epigenetics first neighbours. Differentially regulated genes which are not characterised as epigenetic genes but directly interact with them can be identified within one step networks. Complete lists of differentially regulated interacting partners are provided in Supplementary excel file 5.7. The 'bad' patient prognosis CAM vs. ANM network identifies 153 differentially regulated genes that directly interact with epigenetic genes (Figure 5.15A).

The 'good' patient prognosis CAM vs. ANM network identifies 38 differentially regulated genes, which directly interact with epigenetic genes (Figure 5.15B). In combination, utilizing the interactome provides a list of around 190 differentially regulated genes that directly interact with genes involved epigenetics regulation.

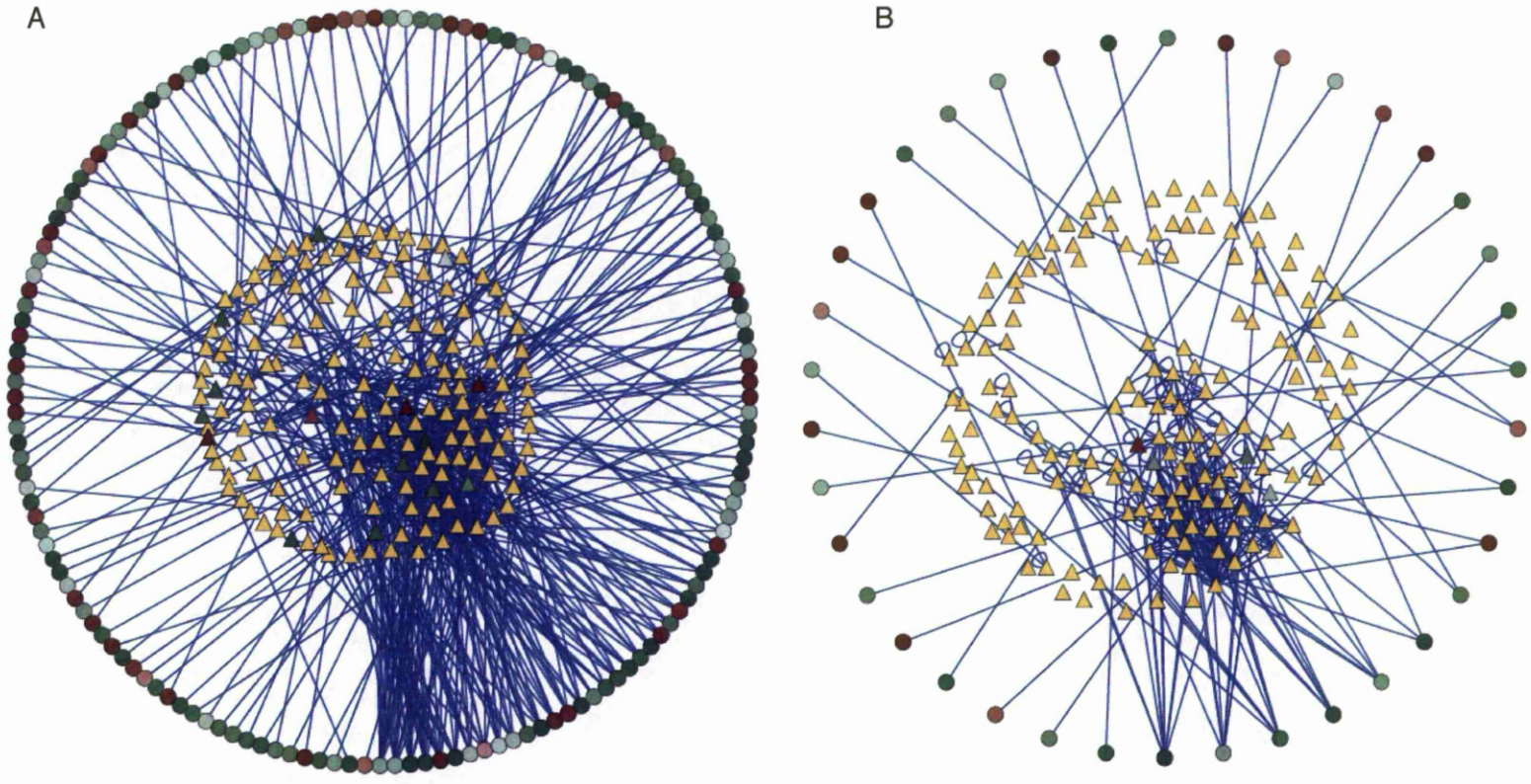



**Figure 5.1**. 164 Epigenetic chromatin modification enzyme and epigenetic chromatin remodelling factors (centre triangular nodes) and one step differentially regulated interacting partners (outer circular nodes). (A) 'Bad' patient prognosis group, CAM vs. ANM. (B) 'Good' patient prognosis group, CAM vs. AN. Nodes represent genes and blue connectors a protein-protein interaction, red nodes represent increased expression and green nodes represent decreased expression. Yellow nodes are un-differentially regulated epigenetic genes.

### 5.1.6 Statistically different gene expression between prognosis groups

As a result of correlation analysis described above it became clear that gene expression patterns of patients with 'bad' and 'good' prognosis scores were different. The 'bad' patients often displayed larger changes in gene expression than patients with 'good' prognosis scores. Therefore, a T-test was applied to identify genes with significantly different gene expression between the 'good' and 'bad' prognosis sub-groups.

Figure 5.16 displays genes with statistically significant difference in log fold-change between 'good' and 'bad' patient sets, at three different thresholds, $p \leq 0.05$, $p \leq 0.01$ and $p \leq 0.005$. At the $p \leq 0.05$ level, 191 genes are classified as having statistically significant different gene expression between 'good' and 'bad' prognosis patients (data supplied in Supplementary excel file 5.8). Seventy-three of which have increased expression in 'bad' prognosis patients and 113 have increased expression in the 'good' prognosis patients. At the $p \leq 0.01$ level, 33 genes are classified as having statistically significant different gene expression between 'good' and 'bad' prognosis patients (Table 5.16). Only nine of which have increased expression in 'bad' prognosis patients and 24 have increased expression in the 'good' patient sub-groups. At the $p \leq 0.005$ level, 16 genes are classified as having significantly different gene expression between 'good' and 'bad' prognosis patients (Table 5.15). Four genes have increased expression in 'bad' prognosis patients and 14 have increased expression in the 'good' prognosis patients.

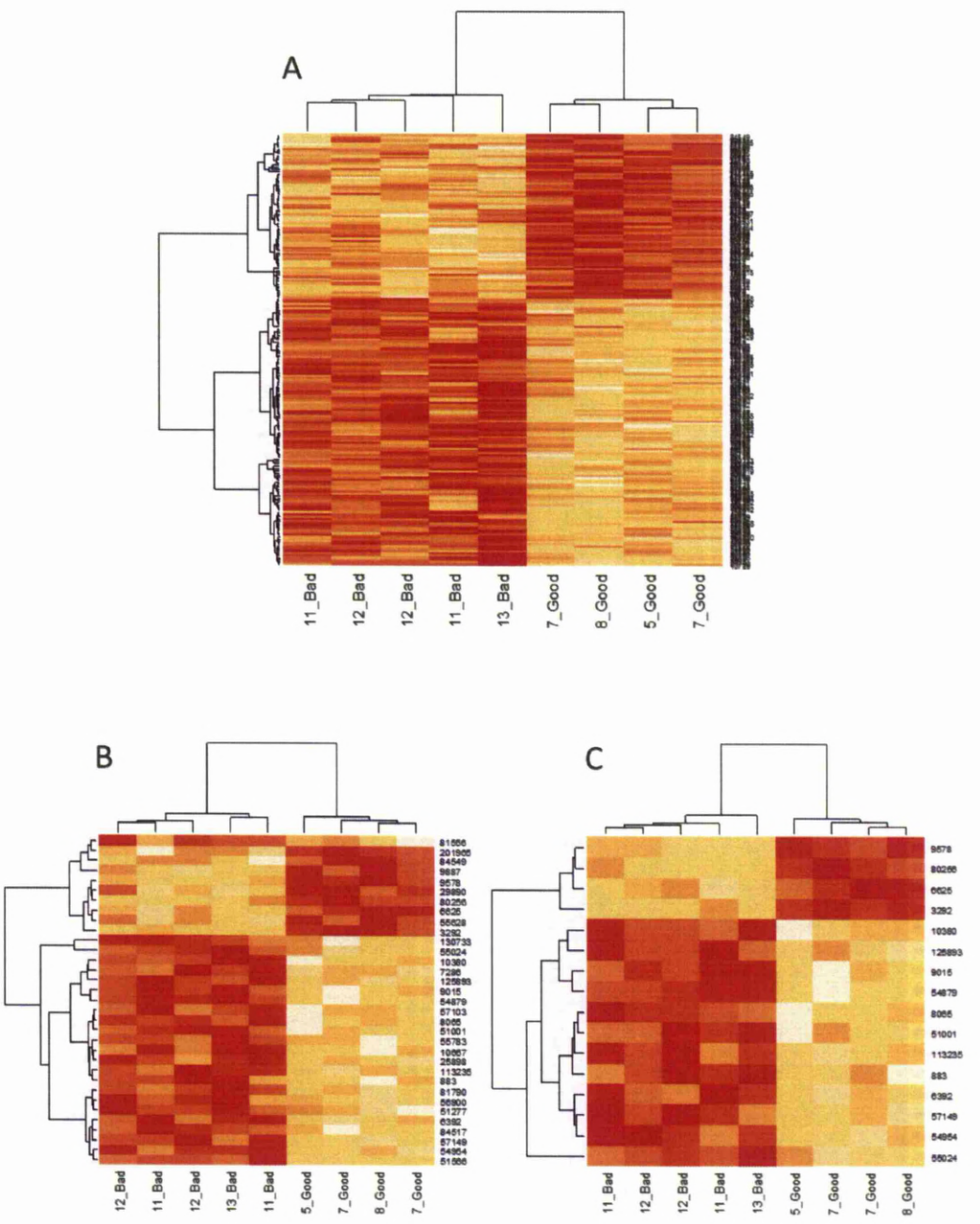



**Figure 5.16**. Genes with statistically significant difference in log fold change between 'good' and 'bad' patient sets. (A) Genes statistically significant p≤0.05. (B) Genes statistically significant p≤0.01 (C) Genes statistically significant p≤0.005. Yellow colouring represents under-expression and red colouring represents over-expression. Patient labels represent patient prognosis score followed by associated 'Good' or 'Bad' prognosis category.

| Statistically different (p≤0.01) | | Statistically different (p≤0.005) | |
|---|---|---|---|
| Gene name | Direction of change | Gene name | Direction of change |
| *C15orf44* | up in 'good' | *CDC42BPB* | up in 'bad' |
| *RWDD4A* | up in 'bad' | *KIAA1539* | up in 'bad' |
| *MAK16* | up in 'bad' | *SNRNP70* | up in 'bad' |
| *SMG7* | up in 'bad' | *Hsd17b1* | up in 'bad' |
| *CDC42BPB* | up in 'bad' | *bpnt1* | up in 'good' |
| *RBM15B* | up in 'bad' | *ZNF816A* | up in 'good' |
| *KIAA1539* | up in 'bad' | *Taf1a* | up in 'good' |
| *SNRNP70* | up in 'bad' | *ST7L* | up in 'good' |
| *ZNF407* | up in 'bad' | *CUL5* | up in 'good' |
| *Hsd17b1* | up in 'bad' | *MTERFD1* | up in 'good' |
| *Tmem178* | up in 'good' | *Slc46a1* | up in 'good' |
| *BANK1* | up in 'good' | *CCBL1* | up in 'good' |
| *bpnt1* | up in 'good' | *LOC100130320* | up in 'good' |
| *Tuft1* | up in 'good' | *lyrm1* | up in 'good' |
| *ZNF816A* | up in 'good' | *FAM120C* | up in 'good' |
| *Taf1a* | up in 'good' | *BANK1* | up in 'good' |
| *ST7L* | up in 'good' | | |
| *C12orf5* | up in 'good' | | |
| *CUL5* | up in 'good' | | |
| *MTERFD1* | up in 'good' | | |
| *FTSJD1* | up in 'good' | | |
| *FARS2* | up in 'good' | | |
| *RCHY1* | up in 'good' | | |
| *Slc46a1* | up in 'good' | | |
| *CCBL1* | up in 'good' | | |
| *Rnf170* | up in 'good' | | |
| *Tmem167b* | up in 'good' | | |
| *DNAJC27* | up in 'good' | | |
| *LOC100130320* | up in 'good' | | |
| *Arpm1* | up in 'good' | | |
| *lyrm1* | up in 'good' | | |
| *FAM120C* | up in 'good' | | |
| *ARMCX3* | up in 'good' | | |

**Table 5.85.** Genes with statistically different gene expression (p≤0.01 and p≤0.005) in 'good' and 'bad' patient prognosis sets, CAM vs. AN.

Previously in section 5.3.3, a statistical analysis was performed on the 'good' CAM vs. ANM dataset and the 'bad' CAM vs. ANM dataset to define which genes were differentially regulated in CAMs compared to ANMs, for 'good' and the 'bad' patient

sub-groups. In theory, genes significantly changed in the 'good' CAM sub-group vs. ANMs may also be statistically different between 'good' and 'bad' patients in the gene lists above. However, this may not always be the case, as small consistent changes may occur and therefore be classed as statistically significant in the 'good' CAM vs. ANM dataset. However, if the same consistent values are seen in the 'bad' CAM vs. ANM dataset, they will not be detected in the analysis of differences between gene expression in 'good' and 'bad' patient subgroups (Figure 5.17).

The number of genes classed as statistically changed in the 'bad' patient prognosis sub-set is far more than those in the 'good' patient prognosis set. Therefore, as expected far more genes, 46 (3.5%) compared to 12 (4%), are statistically changed in the 'bad' CAM vs. ANM dataset and have statistically different gene expression profiles compared to the 'good' patient sub-group.

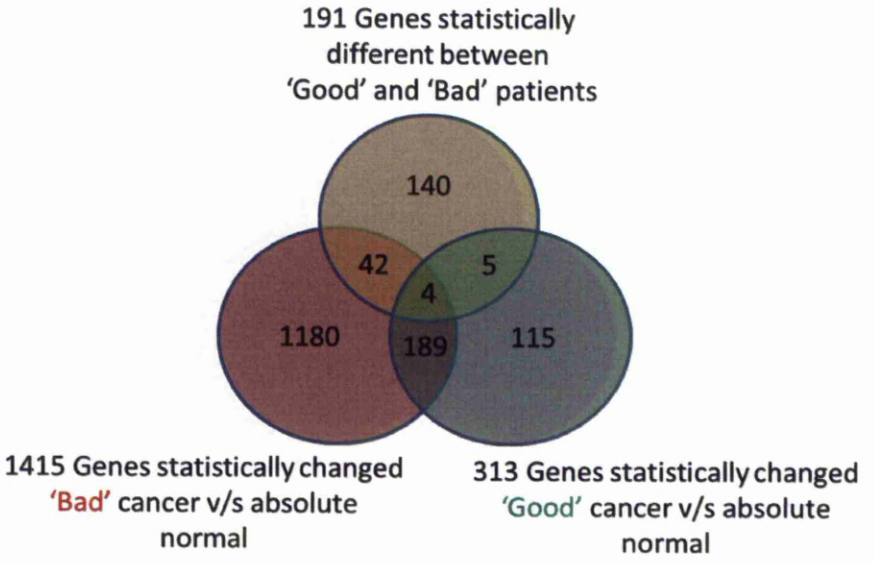



**Figure 5.17**. Venn diagram representing the crossover of genes identified as statistically significant ($p \leq 0.05$) in the 'good' and 'bad' patient CAM vs. ANM datasets to those identified as having statistically different gene expressions ($p \leq 0.05$) between the 'good' and 'bad' patient prognosis groups. Numbers of genes outside the circles represent total numbers within each category.

### 5.1.7 Correspondence analysis

Previously in chapter 4, section 4.3.2.1 multidimensional scaling was used to plot the genes and pathways within 3 dimensional plots based on either their similarities of genes or similarities of pathways. It became apparent that linking clusters of genes within one plot to the associated clusters of pathways within another plot was difficult. Odds ratios were calculated to define pathways that were over-represented in each dataset, using all Reactome™ pathways. Although this has been carried out on all datasets, only the CAM vs. ANM dataset is shown as an example (Figure 5.18). Odds ratios greater than 2 were considered over-represented in all subsequent correspondence analysis.

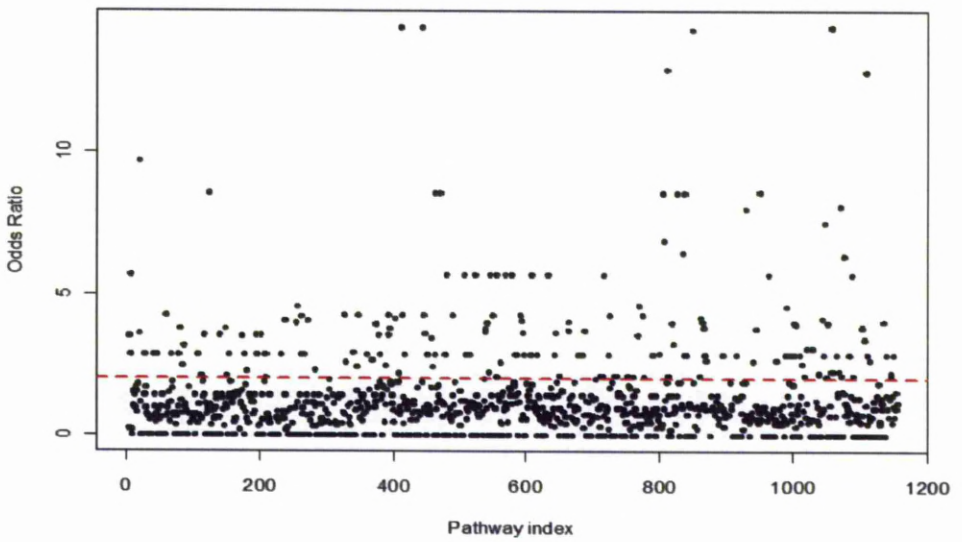
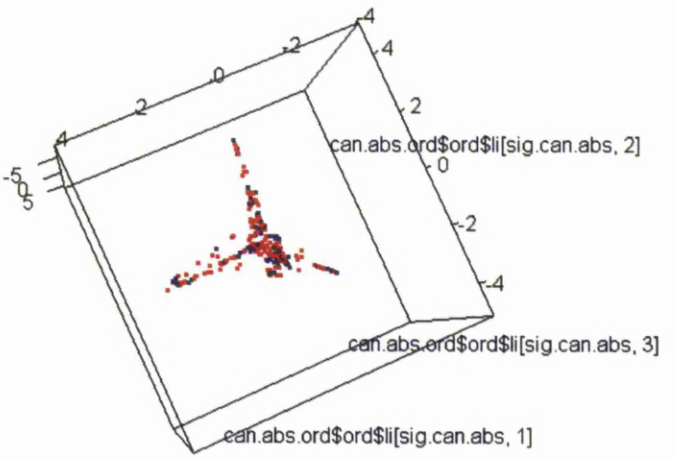



**Figure 5.18.** Odds ratio's for each individual pathway in the Reactome™ database. Red dotted line represents an odds ratio of 2.

Here, using correspondence analysis, clusters of genes and clusters of pathways are plotted together on a single 3 dimensional plot, allowing the easy interpretation of genes affecting pathways and vice versa, as shown below; genes are represented by red dots and pathways are represented by blue dots:




CAM vs. ANM, correspondence analysis of all significantly changed genes and all over-represented pathways are displayed in Figure 5.19. As the positioning of points on the plot represents similarity of pathways and gene membership, those clustered close together must have high similarity. Using the three different dimensions together allows the identification of a dense cluster of pathways and genes, identified by the turquoise box. Similar solitary dense clusters were identified in CAM vs. ATM (Figure 5.20) and ATM vs. ANM (Figure 5.21) datasets. Genes and pathways within dense clusters from the CAM vs. ANM dataset are provided in Supplementary excel file 5.9, CAM vs. ATM dataset provided in Supplementary excel file 5.10 and ATM vs. ANM dataset provided in Supplementary excel file 5.11.

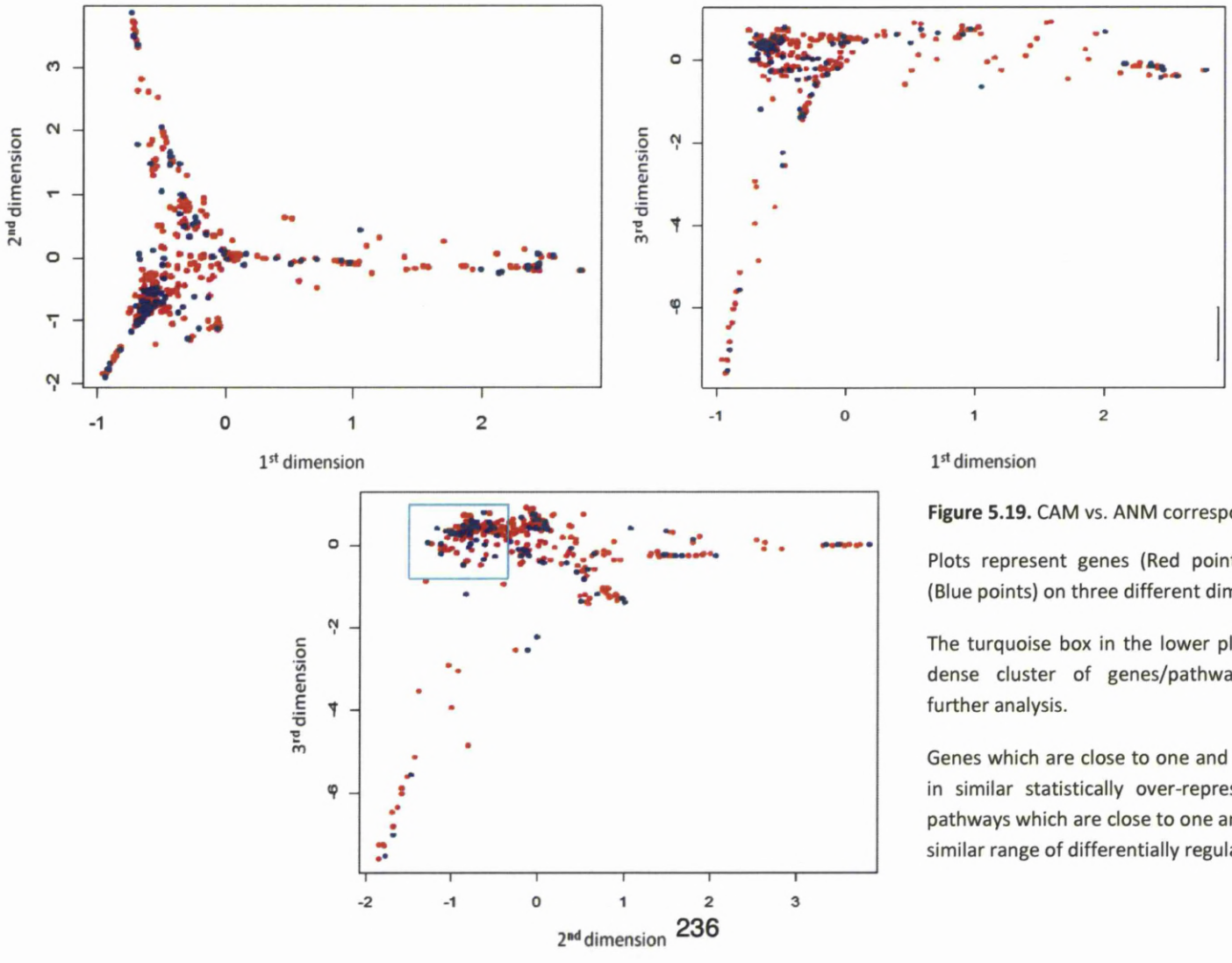



**Figure 5.19.** CAM vs. ANM correspondence analysis

Plots represent genes (Red points) and pathways (Blue points) on three different dimensions.

The turquoise box in the lower plot represents the dense cluster of genes/pathways exported for further analysis.

Genes which are close to one and other are present in similar statistically over-represented pathways, pathways which are close to one and other contain a similar range of differentially regulated genes.

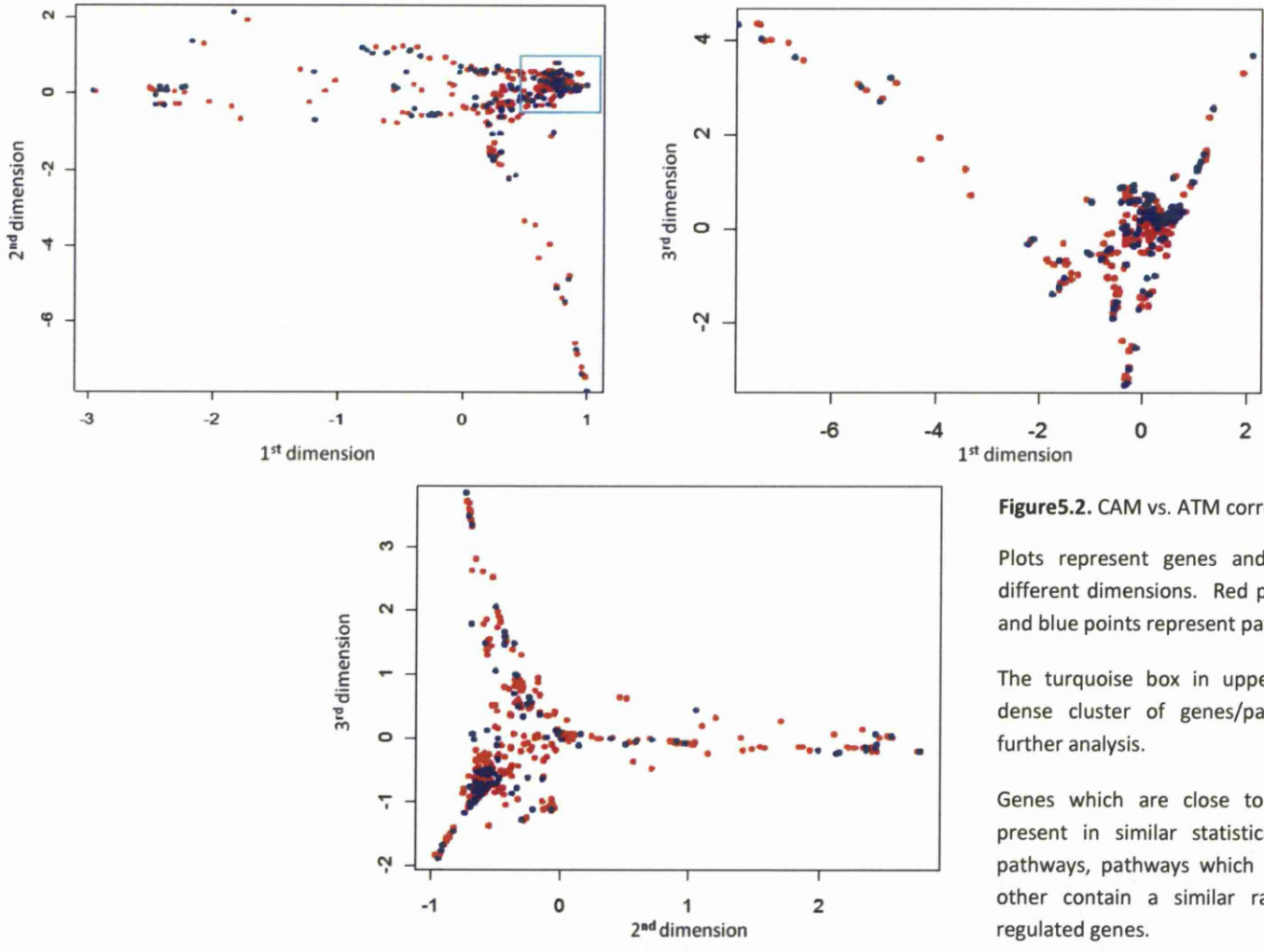



**Figure5.2.** CAM vs. ATM correspondence analysis.

Plots represent genes and pathways on three different dimensions. Red points represent genes and blue points represent pathways.

The turquoise box in upper plot represent the dense cluster of genes/pathways exported for further analysis.

Genes which are close to one and other are present in similar statistically over-represented pathways, pathways which are close to one and other contain a similar range of differentially regulated genes.

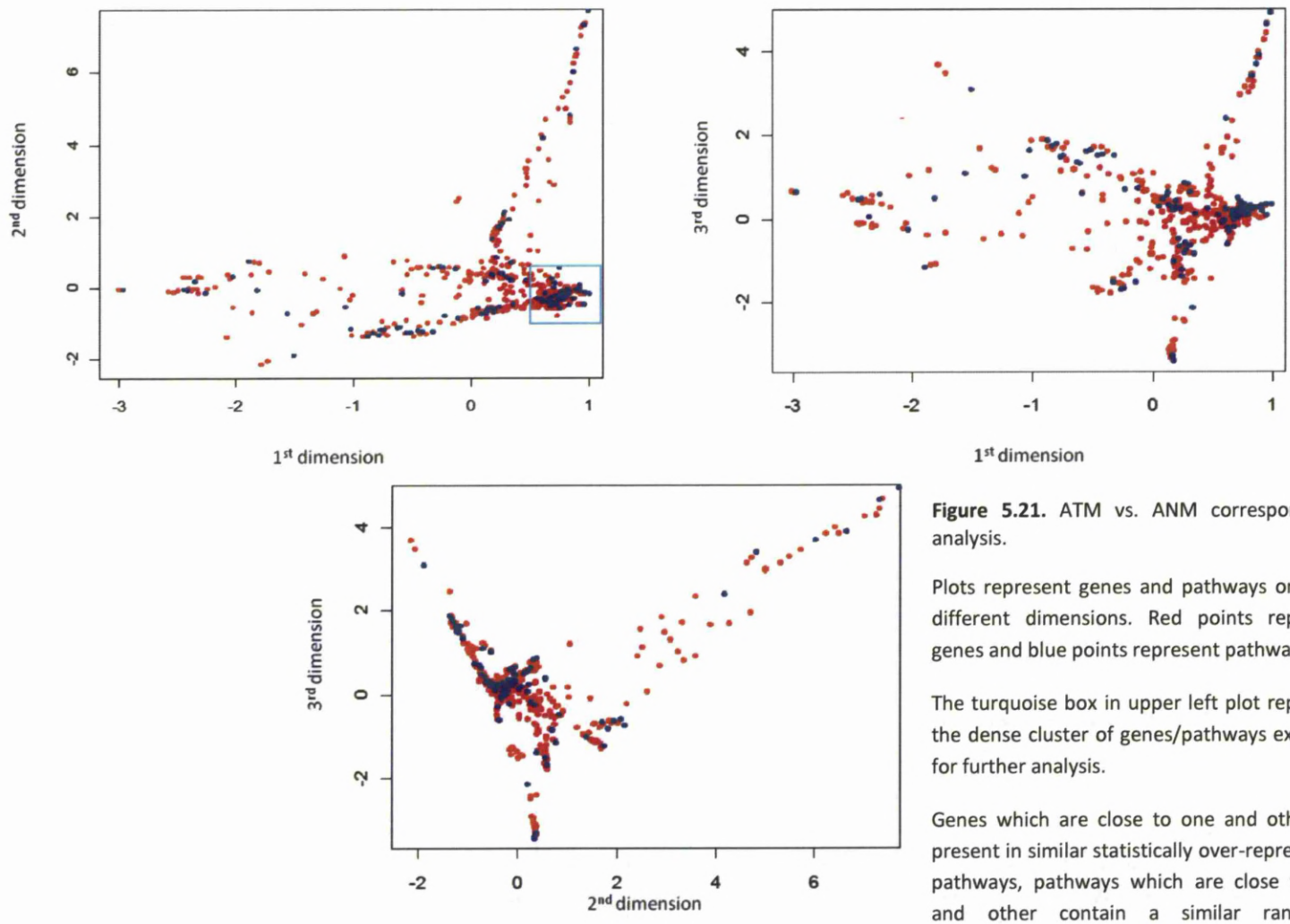



**Figure 5.21.** ATM vs. ANM correspondence analysis.

Plots represent genes and pathways on three different dimensions. Red points represent genes and blue points represent pathways.

The turquoise box in upper left plot represent the dense cluster of genes/pathways exported for further analysis.

Genes which are close to one and other are present in similar statistically over-represented pathways, pathways which are close to one and other contain a similar range of differentially regulated genes.

#### 5.1.7.1 CAM vs. AN

Interferons are glycoproteins that are presented at the cell surface upon the detection of a virus, infection or tumour cells, subsequently triggering an immune response. Specifically, interferon signalling and interferon gamma signalling have many shared genes. Rather than there being a lot of crosstalk between these pathways, the large number of shared genes appears to be because interferon signalling is a parent pathway of interferon gamma signalling. Interferon gamma is released from macrophages and its receptor is present on most cells in the human body, within this dataset the receptor IFNGR2 is up-regulated, along with the majority of differentially regulated genes within the pathways, suggesting a stimulation of interferon genes and an anti-viral state. This increase in interferon activity could be related to interferon's known relation to increased inflammation (Billiau, 1987) as inflammation is known to be precede gastric cancer (Lu et al., 2006).

Many different pathways involving AKT and CREB have been over represented; within the dataset *AKT2* itself is up regulated, whilst *AKT1* and *CREB* are down regulated. In general, *AKT* phosphorylates many different targets in the nucleus, including *CREB*, the phosphorylation of *CREB* leads to its activation and the increased transcription of genes regulating cell survival including the proto-oncogene *c-fos* (Du and Montminy, 1998; Ponti et al., 2002). Over-represented pathways containing these genes include downstream immune signalling, and CAMK IV-mediated phosphorylation of *CREB*. In addition, *PTEN*, which is up regulated, negatively regulates the AKT pathway. The CAMK IV-mediated

phosphorylation of CREB pathway members is a process that occurs in a large range of signalling pathways, such as *FGF* and PDGF signalling and therefore differential regulation of its related genes allows for a wide range of downstream targets (Tan et al., 1996; Wang et al., 2011). The differentially regulated gene list within this dataset are all down regulated.

Three genes involved in *RAF* activation, including the oncogene *KRAS* and *RAF* itself are shown to be down regulated. In its active form, the GTPase *KRAS* converts GTP to GDP, simultaneously activating *c-RAF* (Zhang et al., 1993). This pathway is a downstream component of many signalling pathways, which when differentially regulated, as shown by the many activating mutations of KRAS found in cancers, can have a range of detrimental effects on the cell (Poulogiannis et al., 2012). Within cancers, a mutation of *KRAS* leads to its constant activation, leading to a down-stream propagation signal of constant growth factor stimulation.

The dissociation of glucokinase from its regulatory protein appears to be down regulated, as glucokinase phosphorylates glucose, converting it into glucose-6-phosphate, thus suggests a decrease in glucose catabolism (Iynedjian, 2009). Conversely however, a pathway involving the transport of glucose molecules across the cell membrane is over-represented with two isoforms of the solute transporters, *SLC2A11* and *SLC2A10* being up regulated.

#### 5.1.7.2 CAM vs. ATM

As in the CAM vs. ANM dataset, Interferon signalling and interferon gamma signalling pathways are over-represented, yet this time the just over half of the differentially regulated genes are down-regulated within these pathways. This lack

of distinctive increase or decrease of interferon signalling within cancer cells compared to adjacent cells may represent the similar level of inflammation induced within the cancer and adjacent environment.

Two pathways involving the regulation of apoptosis are over-represented, with pro-survival *BCL2* family members being down regulated and pro-apoptotic proteins being up regulated, overall this should result in an increase in apoptotic cell death. The RAF activation pathway, the RAF phosphorylating MEK pathway and MEK activation pathway are all over-represented, with *KRAS* and *NRAS* over-expressed in all three. These processes take place in a large number of different signalling cascades suggesting *KRAS* and *NRAS* as possible bottlenecks, where by their differential expression could affect a large number of processes. The increase in expression of *KRAS* and *NRAS* suggest an increase in activity of their target *RAF*, which in turn phosphorylates and activates *MEK*, which relates to increased cell growth and division.

#### 5.1.7.3 ATM vs. ANM

Thirty-seven genes are differentially regulated in the innate immune system, with the majority up-regulated, suggesting an increase in immune response, possibly inflicting inflammation and a cancer initiation supportive environment. It is interesting that this is such a largely over-represented pathway with many differentially regulated genes, are occurring in ATMs myofibroblast and inflammation is an early cancer event.

There are 27 genes differentially regulated, in the over-represented 'transmission across chemical synapses' pathway. Ten increased expression and 17 with

decreased expression. As transmission across chemical synapses only occur in neurons, this pathway is not applicable to the myofibroblast, yet interestingly the connection between the cancer microenvironment and neuronal pathways has been made previously. In relation to our previous findings in chapter 4, section 4.2.3.1, we detected differential regulation of a range of metabolic pathways similar to the previously described 'Reverse Warburg effect', whereby fibroblasts carry out aerobic glycolysis to produce lactate, to feed the cancer cells (Pavlides et al., 2009b). Pavlides et al, went on to state that the 'Reverse Warburg effect' and 'neuron-glia metabolic coupling' are analogous biological processes (Magistretti, 2009; Pavlides et al., 2010b).

The over-represented L1CAM interactions pathway contains, 24 differentially regulated genes, three isoforms of Laminin itself, *LAMA1, LAMB1* and *LAMC1*, are over-expressed. Laminins are extracellular glycoproteins which make up part of the basement membrane (Yurchenco, 2011) and interestingly in breast cancer, laminin has been shown to be increased in fibroblasts from the tumour margin, but not in fibroblasts directly from the cancer region (Broes et al., 1988). Thus the difference in laminin expression between fibroblasts from different regions could represent the ability of the cancer to metastasise through the basement membrane.

Overall, interesting over-represented pathways and promiscuous genes identified within the dense clusters for each dataset represent potential points of therapeutic intervention. The alteration of a few or even a single gene will have effects on a large number of over-represented pathways. Using this technique it is possible to

carefully select a gene, by understanding the range of pathways it is involved in and those believed be causative of the cancer or adjacent phenotype.

### 5.1.8 Metabolic signature

As described in the previous chapter the 'Reverse Warburg effect' has been described as a mechanism of metabolic interaction between cancer cells and cells within the cancer microenvironment. To investigate if evidence for similar process could be identified in our data, a targeted analysis was performed to identify changes in metabolic pathways in each dataset. For this study, the Reactome dataset was used without the application of correspondence filters and metabolic pathways were identified as over-represented based on an odds ratio score ≥2 (Table 5.17).

Within the CAM vs. ANM and the CAM vs. ATM dataset, 3 cycles of the fatty acid β-Oxidation pathway were found to be over-represented. This signature was found to be reduced in ATM vs. ANM dataset, with only one cycle of the B-oxidation pathway being over-represented. Overall, CAM vs. ANM dataset reveals the largest number of differentially regulated genes relating to the largest selection of fatty acid B-oxidation pathways. Significantly, the expression of genes within these pathways was generally found to be up in the cancer and adjacent compared to normal samples, however as cancer-associated fibroblasts have a slightly lower up-regulation than the adjacent fibroblasts, the associated genes appear reduced when comparing CAM vs. ATM.

For saturated fatty acids, there are seven cycles through β-oxidation with each cycle of oxidation requiring four individual enzymatic steps. *HADHA* and *HADHB* are

differentially expressed in all three datasets, these subunits form the tri-functional protein; hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydralase, which is involved in the three last enzymatic steps of β-Oxidation. The enzyme required for the first enzymatic step is not so uniformly expressed, with different members of the acyl-CoA dehydrogenase *(ACAD)* family being differentially expressed within the different datasets. The different *ACAD* members have not been annotated within the Reactome pathways and so had to be identified within the differentially regulated gene sets. For the CAM vs. ANM *ACAD9* was differentially regulated, for the ATM vs. ANM dataset *ACAD9* and *ACAD11* were differentially regulated and for the CAM vs. ATM dataset *ACADM* member was differentially regulated. Each cycle of β-oxidation produces one molecule of acetyl-CoA and an extra acetyl-CoA generated from the last cycle, with the number of cycles needed for a single fatty acid being represented by ($C_{2n}$) n - 1 oxidations, therefore for palmitate ($C_{16}$, $n = 8$) a maximum of seven passes is required.

The classical 'Warburg effect' describes a process whereby surrounding fibroblasts upon activation by cancer cells carry out aerobic glycolysis, producing lactate and pyruvate which can then be transported back into nearby cancer cells, via membrane transporters, before entering the TCA cycle to drive for oxidative phosphorylation (Pavlides et al., 2010b; Pavlides et al., 2009b). However here we have not identified an alteration in such glycolytic pathways that would result in an end product of pyruvate or lactate but have identified an increase in fatty acid β-oxidation and production of acetyl coA. Acetyl CoA is a high energy source, with each molecule of acetyl CoA producing 10 molecules of ATP, when fed into the

citric acid cycle. Whilst acetyl CoA itself cannot be transported between cells, it can be converted into ketone bodies such as acetoacetate or 3-hydroxybutrate, which are known to 'fuel' tumour cell metabolism when converted from pyruvate (Pavlides et al., 2010a).

Studies have shown that the *MCT4* transporter is up regulated in cancer-associated fibroblasts displaying the 'Reverse Warburg' phenotype (Pavlides et al., 2009b). As this transporter can also transport ketone bodies out of the cell datasets were screened for differential regulation of related transporters (Table 5.16). In this study, the same transporter *SLC1A3 (MCT4)* was found to be progressively up-regulated based on the proximity of the fibroblasts to the cancer cells, being up-regulated 5.5 fold in CAMs and 3.5 fold in ATMs.

In addition, 4 glucose transporters that are alternative transporters to those shown to be involved in the 'Reverse Warburg effect' in breast cancer (Pavlides et al., 2009b), which were; *Slc2a6* solute carrier family 2 *(GLUT6), Slc2a5* solute carrier family 2 *(GLUT5), Slc2a3* solute carrier family 2 (GLUT3) and *Slc2a8* solute carrier family 2 *(GLUT8)* were all found to be up-regulated in this study.

| CAM vs. ANM | | |
|---|---|---|
| **Genename** | **Over-represented pathway** | **Fold change** |
| *HADHA* | Beta oxidation of palmitoyl-CoA to myristoyl-CoA | 1.435057531 |
| *HADHB* | Beta oxidation of palmitoyl-CoA to myristoyl-CoA | 1.245247808 |
| *HADHA* | Beta oxidation of lauroyl-CoA to decanoyl-CoA-CoA | 1.435057531 |
| *HADHB* | Beta oxidation of lauroyl-CoA to decanoyl-CoA-CoA | 1.245247808 |
| *HADHA* | Beta oxidation of hexanoyl-CoA to butanoyl-CoA | 1.435057531 |
| *HADHB* | Beta oxidation of hexanoyl-CoA to butanoyl-CoA | 1.245247808 |
| *CRAT* | Beta-oxidation of very long chain fatty acids | 1.440261377 |
| *ACOX1* | Beta-oxidation of very long chain fatty acids | 1.354181826 |
| *ACAA1* | Beta-oxidation of very long chain fatty acids | 1.260756855 |
| *HADHA* | mitochondrial fatty acid beta-oxidation of unsaturated fatty acids | 1.435057531 |
| *DECR1* | mitochondrial fatty acid beta-oxidation of unsaturated fatty acids | 1.26565037 |
| *HADHB* | mitochondrial fatty acid beta-oxidation of unsaturated fatty acids | 1.245247808 |
| CAM vs. ATM | | |
| **Genename** | **Over-represented pathway** | **Fold change** |
| *HADH* | Beta oxidation of hexanoyl-CoA to butanoyl-CoA | -1.237411811 |
| *HADHA* | Beta oxidation of hexanoyl-CoA to butanoyl-CoA | -1.14483521 |
| *HADH* | Beta oxidation of lauroyl-CoA to decanoyl-CoA-CoA | -1.237411811 |
| *HADHA* | Beta oxidation of lauroyl-CoA to decanoyl-CoA-CoA | -1.14483521 |
| *HADHA* | Beta oxidation of palmitoyl-CoA to myristoyl-CoA | -1.14483521 |
| *HADHA* | mitochondrial fatty acid beta-oxidation of unsaturated fatty acids | -1.14483521 |
| *ACADM* | mitochondrial fatty acid beta-oxidation of unsaturated fatty acids | 1.147683425 |
| ATM vs. ANM | | |
| **Genename** | **Over-represented pathway** | **Fold change** |
| *HADHB* | Beta oxidation of myristoyl-CoA to lauroyl-CoA | 1.248679319 |
| *HADHA* | Beta oxidation of myristoyl-CoA to lauroyl-CoA | 1.514652933 |
| *ACAA1* | Beta-oxidation of very long chain fatty acids | 1.348725926 |
| *CRAT* | Beta-oxidation of very long chain fatty acids | 1.572007723 |
| *SC5DL* | Cholesterol biosynthesis | 1.344902646 |
| *NSDHL* | Cholesterol biosynthesis | 1.464412774 |
| *DHCR24* | Cholesterol biosynthesis | 1.491199752 |
| *SQLE* | Cholesterol biosynthesis | 1.497376826 |
| *FDPS* | Cholesterol biosynthesis | 1.518657587 |
| *HMGCR* | Cholesterol biosynthesis | 1.643579611 |
| *CYP51A1* | Cholesterol biosynthesis | 1.713404237 |
| *FDFT1* | Cholesterol biosynthesis | 1.713556629 |
| *SC4MOL* | Cholesterol biosynthesis | 2.027605686 |
| *DHCR7* | Cholesterol biosynthesis | 2.086732666 |
| *IDI1* | Cholesterol biosynthesis | 2.145916269 |
| *TM7SF2* | Cholesterol biosynthesis | 2.416819435 |
| *HMGCS1* | Cholesterol biosynthesis | 4.403840803 |
| *LSS* | Cholesterol biosynthesis | 6.094901379 |
| *AGPAT5* | Triglyceride Biosynthesis | -1.700541207 |
| *ACSL1* | Triglyceride Biosynthesis | 1.186610321 |
| *ACSL3* | Triglyceride Biosynthesis | 1.214242907 |
| *HSD17B12* | Triglyceride Biosynthesis | 1.231835938 |
| *ELOVL5* | Triglyceride Biosynthesis | 1.291191624 |
| *AGPAT3* | Triglyceride Biosynthesis | 1.431333726 |
| *AGPAT2* | Triglyceride Biosynthesis | 1.437392973 |
| *LPIN1* | Triglyceride Biosynthesis | 2.79518078 |

**Table 5.96.** Reactome metabolic changes, over-represented fatty acid pathways (odds ratio ≥2) and differentially regulated genes ($p \leq 0.05$), for the CAM vs. ANM, CAM vs. ATM and ATM vs. ANM datasets.

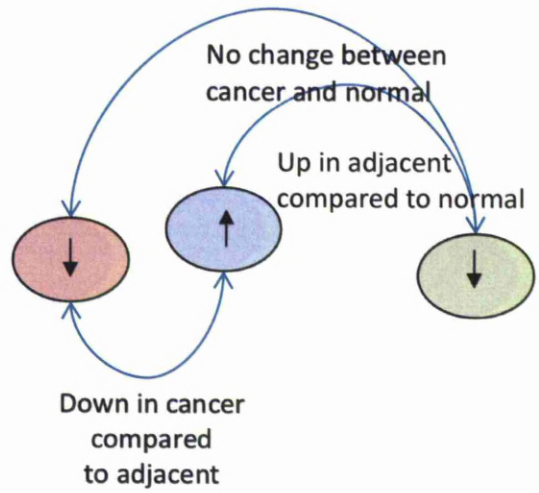
Interestingly the fatty acid transporter, *SCL27A1* was also found to be up-regulated in CAMs vs. ANMs and the CAM vs. ATM fibroblasts, but not in ATM vs. ANM fibroblasts, which supports our earlier findings of an over-representation of fatty acid-oxidation pathways within cancer associated fibroblasts.

| CAM vs. ANM | | |
|---|---|---|
| Genename | Over-represented pathway | Fold change |
| *SLC2A11* | Facilitative Na+-independent glucose transporters | 1.756220546 |
| *SLC2A10* | Facilitative Na+-independent glucose transporters | 1.418203451 |
| *SLC16A3* | Bile salt and organic anion SLC transporters | 5.435000361 |
| *SLC16A7* | Bile salt and organic anion SLC transporters | -1.498706699 |
| *SLC27A1* | Transport of fatty acids | 1.414045494 |
| CAM vs. ATM | | |
| Genename | Over-represented pathway | Fold change |
| *SLC16A7* | Bile salt and organic anion SLC transporters | -1.149752179 |
| *SLC16A3* | Bile salt and organic anion SLC transporters | 1.527248631 |
| *SLC27A4* | Transport of fatty acids | 1.31118933 |
| ATM vs. ANM | | |
| Genename | Over-represented pathway | Fold change |
| *SLC16A1* | Bile salt and organic anion SLC transporters | -1.488331374 |
| *SLC16A3* | Bile salt and organic anion SLC transporters | 3.591340078 |

**Table 5.17.** Over-represented transporter pathways (odds ratios >2) identified within the Reactome database and associated with differentially regulated transporters ($p \leq 0.05$), for the CAM vs. ANM, CAM vs. ATM and ATM vs. ANM datasets.

Cholesterol and triglyceride biosynthesis was also found to be up-regulated in ATM (Table 5.17), which was confirmed within Metacores™ GeneGo pathways (Figure 5.22). Principal component analysis revealed a 1.6 fold change cut-off to be optimum to reveal changes in the CAM vs. ATM dataset, and was therefore applied throughout this chapter. However, it is interesting to mention that removing this fold change cut-off, yet retaining a p-value ≤0.05 and subsequent Metacore GeneGo pathway analysis reveals cholesterol biosynthesis as a significantly over-represented pathway (Figure 5.22). No change in cholesterol biosynthesis is

detected upon comparison of CAM vs. ANM, within the ATM vs. ANM dataset, differentially regulated genes are up regulated and within the CAM vs. ANM dataset differentially regulated genes are down-regulated. These findings are summarised in the diagram below:




Cholesterol biosynthesis is an alternative route for acetyl-CoA production, and this apparent increase in fatty acid storage as opposed to energy metabolism may indicate an important difference between cancer-associated fibroblasts and adjacent fibroblasts.

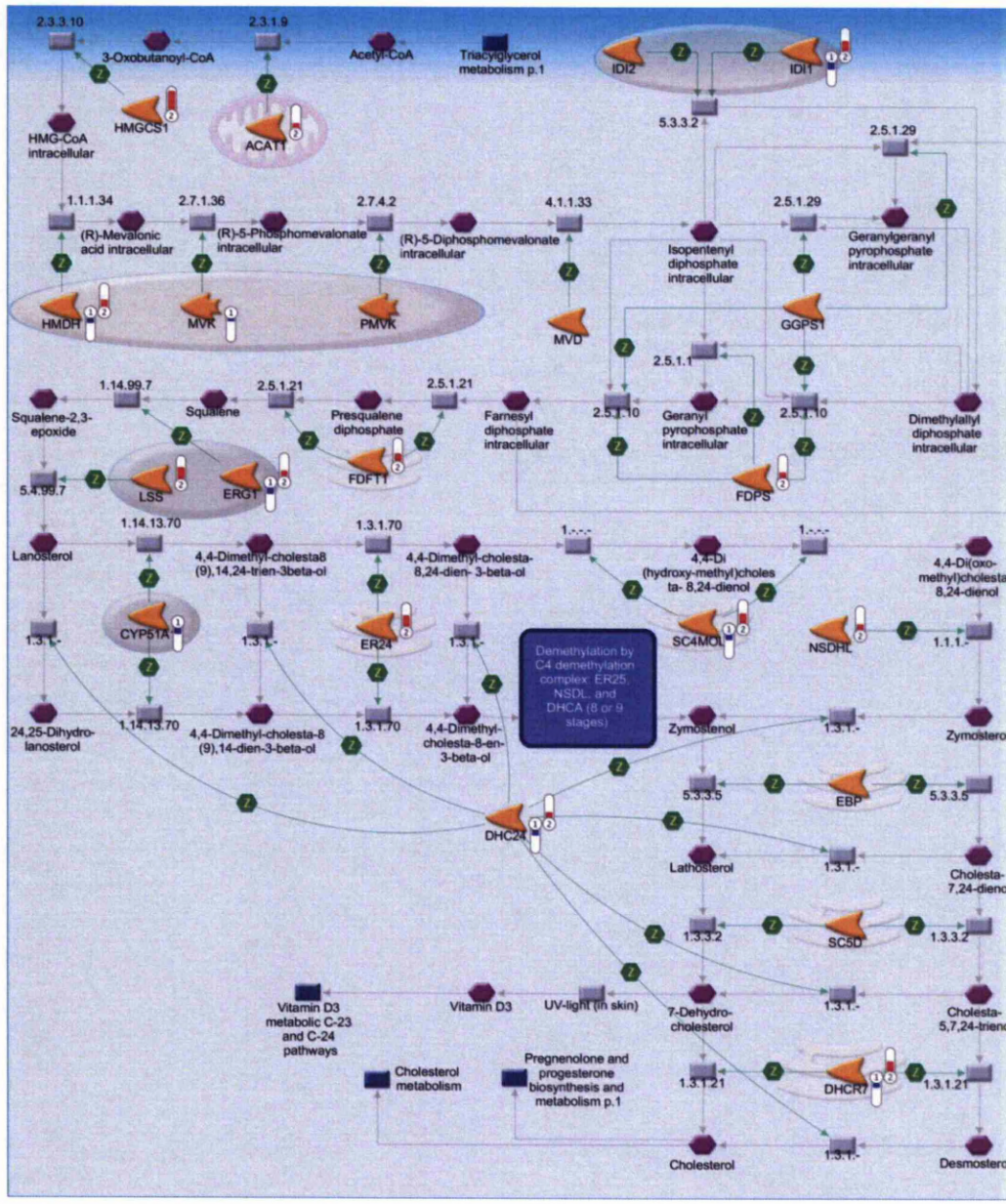



**Figure 5.22.** Metacore Cholesterol Biosynthesis GeneGo pathway. Differentially regulated genes are represented by numbered thermometers within the (1) CAM vs. ATM or (2) ATM vs. ANM datasets. Red thermometers represent up-regulated genes and blue thermometers represent down-regulated genes, with the relative amounts of red/blue colouring representing the size of the change. Orange shapes represent generic enzymes, grey rectangles are reactions and purple hexagons are compounds.

The final data required to support this hypothesis, was evidence of the conversion of acetyl-CoA into a transportable ketone body, such as acetoacetate or 3-Hydroxybutrate. Metacore™ GeneGo pathway analysis revealed the over-represented pathway, 'ketone body biosynthesis in the CAM vs. ANM dataset (Figure 5.23). Importantly the two enzymes involved in important steps leading to the production of acetoacetate are up-regulated within cancer associated fibroblasts.

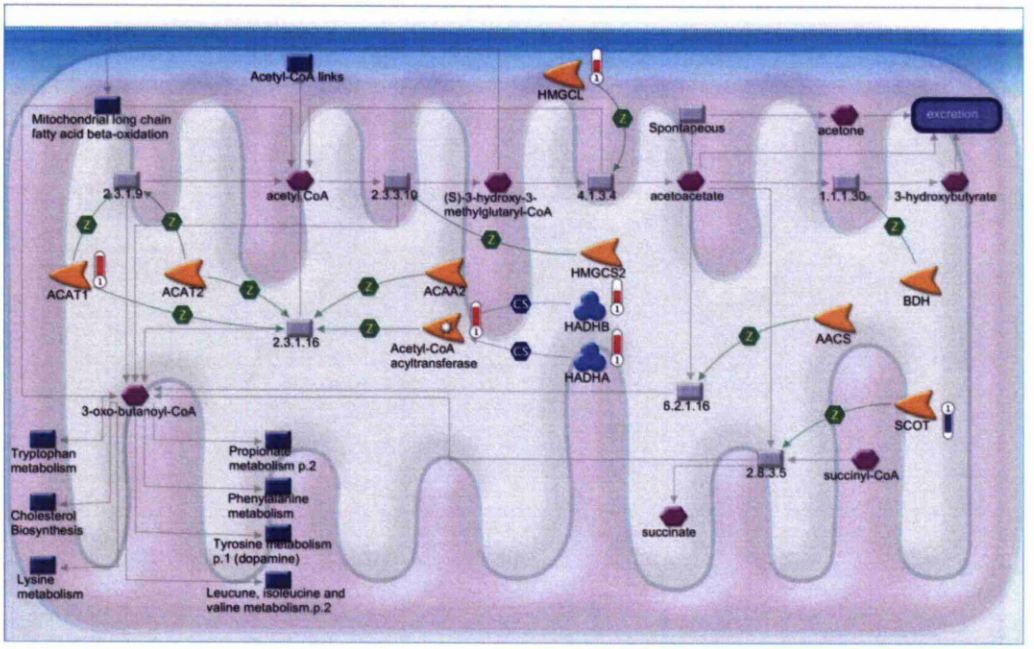



**Figure 5.23.** Metacore Ketone Body synthesis GeneGo pathway. Differentially regulated genes are represented by numbered thermometers within the 1) CAM vs. ANM dataset. Red thermometers represent up-regulated genes and blue thermometers represent down-regulated genes, with the relative amounts of red/blue colouring representing the size of the change. Orange shapes represent generic enzymes, grey rectangles are reactions and purple hexagons are compounds.

## 5.2 Discussion

PCA analysis alone revealed little difference between CAMs and ATMs, however by applying more stringent thresholds than those needed for CAM vs. ANM or ATM vs. ANM comparisons the two myofibroblast populations could be resolved. However, the challenge remained, to establish an appropriate fold change cut-off to apply, whilst retaining the maximum number of genes to allow meaningful pathway enrichment and correspondence analyses. A 1.6 fold change cut-off provided a reasonable segregation of samples and a useable size gene list for further analysis. To be cautious, a 2-fold change gene list pathway analysis, was also performed to provide confidence in 1.6-fold change pathway findings.

Availability of new survival scores during the course of these studies, demonstrated that prognostic scores and survival scores do not always correlate, therefore a decision needed to be made of the most appropriate scoring system to apply to categorise changes related to different stages of tumour development. Plotting the variance of the individual patients fold changes revealed an interesting trend, the larger the variance of the data the worse the tumour classification score. Patients with contradicting prognosis and survival scores, whose outcome is potentially hard to predict, displayed variances, which directly correlated with prognosis, based on tumour stage score and not with survival.

Prognosis (Tumour stage) scores were then used to perform a correlation analysis, comparing individual gene fold changes to prognostic scores. The correlation analysis successfully differentiated those patients with conflicting prognosis and survival scores. Two lists of highly correlated genes, which distinguish between

'good' and 'bad' patient sub-groups show interesting trends, which may provide insights into the processes which operate at different stages of tumour development. Verification of the observed differences is now being performed by rtPCR and/or western blotting.

Throughout this study, several different approaches have been used to provide new insight into the molecular processes that are altered and therefore contribute to the different phenotypic properties of CAMs and ATMs relative to ANMs.

Biological pathways found to be important in CAMs include a large number of cell-cycle regulation pathways, of which all differentially regulated genes are down-regulated, suggesting a decrease in cell cycle progression and cell proliferation. This same trend was observed in ATMs, although fewer pathway members were affected in these cells. Another similar trend in CAMs and ATMs includes DNA damage induced apoptosis or cell cycle checkpoint arrest, involving important caretaker genes, such as BRCA1, which was down regulated in both CAMs and ATMs. Ketone body synthesis was up-regulated in CAMs, whilst in ATMs, the top Metacore™ pathway was cholesterol biosynthesis, with all of the differentially regulated genes being over-expressed. As both pathways utilise acetyl-CoA, this apparent difference within CAMs and ATMs was interesting and was investigated further.

These pathways represent the novel differences between CAMs and ATMs, in all stages of cancer, whereas Metacore™ GeneGo pathway analysis of 'good' and 'bad' patient prognosis CAM vs. ANM represents over-represented biological processes or pathways seen in cancer myofibroblasts at different stages of the disease. This

may be an interesting resource, which could be used as a starting point, to highlight a range of pathways that are differentially active in CAMs and ATMs in different patients.

Biological pathways important in 'bad' patient prognosis CAMs include cell-cycle pathways, suggesting an overall decrease in cell proliferation. Interestingly, three separate over-represented pathways are involved in DNA damage detection and down-regulation of caretaker genes, such as BRCA1. Significantly, this signature is not detectable in 'good' prognosis dataset. As also shown in the CAM vs. ANM dataset, mitochondrial ketone body biosynthesis is over-represented. With enzymes converting acetyl co-A into the soluble ketone, acetoacetate, displaying increased expression. In 'good' patient prognosis cancer myofibroblasts, the top pathways include cell cycle related pathways with many down-regulated genes, and increased expression of a range of anti-oxidant enzymes, whose increase in expression is often indicative of exposure to oxidative stress. Patients with this antioxidant signature would possibly be more resistant to the oxidative stress, which is known to be created by rapidly growing tumours.

Overall, changes in transcription factor gene expression were not large across any of the datasets, but differentially regulated transcription factors did include BRCA1, which was is down-regulated in both CAMs and ATMs. Transcription factors displaying this progressive change include, E2F1, PA2G4, RBL1 and STAT2. In breast cancer low E2F levels are associated with a 'good' outcome (Vuaroqueaux et al., 2007). The only transcription factors displaying this progressive change, with over-representation is STAT2, which is known to contribute to cancer progression

(Gamero et al., 2010). The 27 transcription factors differentially regulated, solely in patients with 'bad' prognosis scores, are currently being investigated by rtPCR may be potentially interesting. In future it would be interesting to investigate if genes regulated by differentially expressed transcription factor are components of pathways which are predicted to contribute to CAM or ATM related phenotypes.

Interestingly we observed that a larger number of genes involved in epigenetic regulation were found to be differentially regulated in the 'bad', compared to the 'good' patient prognosis group, therefore suggesting that epigenetic changes may be more extensive in later stages of gastric cancer. Interestingly ING3 and SMARCA4 were found to be differentially regulated in our study, ING3's family members have previously been linked to poor patient cancer prognosis, and as ING3 is uniquely differentially regulated within the 'bad' patient prognosis cancer associated myofibroblasts this observation merits further investigation. Equally, SMARCA4 has been linked to cancer progression, through effects on the extracellular matrix, suggesting that these changes may contribute to the known role of CAMs in modifying the stromal extracellular matrix.

Within the previous results chapter the idea of the 'Reverse Warburg effect' was introduced it was speculated that a similar process could be taking place within gastric CAMs. Interestingly it seems that gastric CAMs may operate a variation on this process. The classical 'Reverse Warburg effect' describes a mechanism whereby surrounding fibroblasts are programed to carry out aerobic gylcolysis, producing lactate and pyruvate, which can then be transported to nearby cancer cells, via

membrane transporters, entering the TCA cycle for oxidative phosphorylation (Pavlides et al., 2010b; Pavlides et al., 2009b).

In this study we have not identified any alterations in glycolytic pathways that would result in an end product of pyruvate or lactate. However, gastric CAMs have increased expression of the fatty acid transporter SCL27A1 and fatty acid β-Oxidation pathways, with the fatty acid transporter un-differentially expressed and the fatty acid β-Oxidation pathway signature present but much reduced within the ATM vs. ANM dataset. Cycles of β-oxidation produce high energy molecules of acetyl Co-A, with each molecule of acetyl CoA producing 10 molecules of ATP, when feed into the citric acid cycle.

Cholesterol and triglyceride biosynthesis was also found to be up-regulated in ATMs. However, in this case it appears that this may be diverted to drive an increase in fatty acid storage, as opposed to energy metabolism. This may represent an important and novel difference between CAMs and ATMs. As mentioned later cancer cells are thought to induce oxidative stress in surrounding fibroblasts, causing them to become 'activated' and to change their metabolic behaviour.

In comparison, CAMs exhibit changes, which indicate that in these cells acetyl-CoA is converted into a transportable ketone body, acetoacetate, which has been shown to has been shown to 'fuel' tumour cell metabolism (Pavlides et al., 2010a). Bonuccelli (2010) showed that the end products of glycolysis, 3-hydroxy-butyrate and L-lactate stimulate tumour growth. In addition, they show that mitochondrial respiration in increased in cancer cells relative to adjacent stroma cells (Bonuccelli

et al., 2010a). Overall, addition of energy metabolites, ketones and lactate increased cell 'stemnesss' and are shown to correlate with poor clinical outcome, specifically metastasis and re-occurrence. In addition to the well-known role of acetyl-CoA in driving the TCA cycle, acetyl-CoA and ketone bodies have also been shown to increase histone acetylation and drive epigenetic changes (Martinez-Outschoorn et al., 2011). Again it would be interesting to see if increased levels of acetyl-CoA and ketone bodies could drive epigenetic changes in CAMs or ATMs, which in anyway contribute to their altered phenotypes relative to ANMs.

The MCT4 transporter is known to be up regulated in cancer-associated fibroblasts displaying the 'Reverse Warburg' phenotype (Pavlides et al., 2009b); In breast cancer, increased MCT4 is observed in cancer associated fibroblasts with 'bad' patient survival scores and therefore MCT inhibitors were suggested as possible anti-cancer therapies (Bonuccelli et al., 2010a). Due to its strong links with the 'Reverse Warburg effect', we looked into the MCT4 transporter family and found they can also transport ketone bodies out of the cell. Significantly MCT4 appears to be progressively up-regulated in gastric myofibroblasts (CAM>ATM>ANM), being up-regulated 5.5 fold in CAMs and 3.5 fold in ATMs. A model of gastric myofibroblast metabolic remodelling is presented in chapter 6, Figure 6.1.

Finally, we were interested to know if hypoxia or oxidative stress may contribute to the 'metabolic reprogramming' of gastric myofibroblasts. There is a growing body of evidence relating cancer cell induced 'aerobic glycolysis' and mitophagy. Hypoxia is thought to be caused by the cancer cells, increasing or activating the transcription factor HIF within the cancer associated fibroblasts. HIF binding to hypoxia response

element (HRE) sites within the MCT4 promoter region, directly causes an increase in MCT4 expression. The abolishment of HIF or use of an oxidative stress inhibitor removed hypoxia induced promotion of MCT4 (Robey et al., 2005; Ullah et al., 2006; Whitaker-Menezes et al., 2011). The deletion of the oxidative defence gene, Jun D, in cancer stroma of mice resulted in an increased activated cancer associated fibroblast and increased metastasis and angiogenesis (Toullec et al., 2010). In light of these observations it will be important to investigate if oxidative stress can trigger the observed increase in 'fatty acid metabolism' in gastric myofibroblasts, or drive ATMs towards a more dangerous CAM phenotype. As ketone bodies produce more energy than lactate and also use less oxygen (Cahill, Jr. and Veech, 2003) they are a more powerful metabolite for cancer cells. These finding suggest that diabetics high ketonic diets may need to be re-thought to improve prognosis in these patients (Bonuccelli et al., 2010b). Work in now underway to verify the predicted changes in metabolic status and expression of transport channels in all available CAM and ATM cell lines.

# 6 Chapter 6 Concluding summary

There is now strong evidence that tumours do not simply grow independently in tissues. The growth, proliferation and spread of cancer cells is intimately linked to the microenvironment that surrounds the growing tumour. Fragmented evidence is emerging to show that normal tissue stroma can be inhibitory for the growth and proliferation of cancer cells (Iacopino et al., 2012), however, under conditions of oxidative stress or chronic inflammation, stromal cells become reprogrammed to provide signals and nutrients which aid tissue repair processes. In these, conditions, cells that would normally work to maintain normal tissue function actually aid tumour growth. This process sets up a self-perpetuating cycle of paracrine communication in which cancer cells reprogram stromal cells to produce growth factors, metalo-proteases (MMPs), and nutrients, which together drive further tumour growth, and induce cancer cell migration. The stroma that surrounds a developing tumour may vary in different tissues, or different stages of tumour development. However, there is strong evidence that myofibroblasts for a key part of the tumour stroma in several tissues. To date, most work in this area has been focused on breast, prostate & colon cancers (Barron and Rowley, 2012; Untergasser et al., 2005; Yazhou et al., 2004). From these studies it has become clear that myofibroblasts within the tumour stroma play an important role in all aspects of tumour growth including proliferation and metastasis. However, many of the details relating to the origins of myofibroblasts, the ways in which they become reprogrammed still remain unclear. In addition, it is also not clear to what extent processes observed in one type of tumour niche operate in other tissues, or if

common affects are elicited by the same molecular changes in different tumours, or in different individuals. For example, different components of the same pathway may be changed in different situations resulting in the same net effect. In terms of gastric cancer very few of these questions have been addressed.

In this study we aimed to analyses the global gene expression profiles of a series of primary human gastric myofibroblasts, purified from tissue samples isolated either from the site of a tumour, from matched tissue adjacent to the tumour (from the same patient), or from normal gastric tissues derived from organ downers. Purification of myofibroblast cells and processing of samples for microarray was performed in the lab of Professor. Andrea Varro (University of Liverpool). At this stage data was provided as Mas5 normalised data, which had been subject to in-house quality control. At the time these studies were performed Mas5 was a more widely used form of data normalisation and a method that had been used in many published studies. In a retrospective analysis of this data (described in Chapter 3) we used the AQM package to re-assess overall data quality and provide insight into the possible effects that batch processing may have had on data quality or interpretation. In addition, we also performed a second retrospective study to assess the possible limitations of using the Mas5 normalisation method, as opposed to the now more commonly used RMA method. Reassuringly the AQM analysis of data quality showed that most arrays were of acceptable quality and despite the high number of batch processing dates used in the study, only two arrays were found to be significant outliers, when compared to all other arrays. Also, these two arrays did not appear as outliers, after batch or patient correction was performed.

With respect to the comparisons of using RMA or Mas5 methods to normalise primary data; Relatively small differences in gene numbers or altered processes were detected when using identical stringent thresholds ($p \leq 0.005$, with a $\geq 2$ fold change threshold), similar to those used in our initial analysis (Chapters 3 and 5). However, the potential value of the RMA method does become apparent when using less stringent thresholds, where far more changed genes are represented. In light of this observation it is reasonable to conclude that RMA normalised data may provide a greater range of information than the Mas5 data used in our initial studies. This may be particularly true for the types of multivariate/correspondence analyses described in Chapters 4 & 5. In light of this data we have now initiated further studies to assess the benefits of using RMA normalised data in these approaches. Despite the fact that Mas5 normalisation methods may limit the amount of information available for analysis there is no reason to believe that robust signatures that were identified from this data are not real or significant. Indeed, several of the predicted signatures and functional predictions made from this data have since been verified in the Sanderson lab, including the differential expression of surface transport channels, the altered metabolic status of CAMs and most significantly the fact that myofibroblasts from different prognostic patient subgroups (identified in Chapter 5) actually induce subgroup specific changes in gene expression profiles when AGS gastric cancer cells are conditioned with media derived from CAMs from different prognostic sub-groups. Most significantly, this also applied to detectable differences between the two different 'bad' CAM subgroups. Due to the relatively small number of available primary myofibroblast cell lines available for use in this study, we cannot say with certainty that these are

generic trends that would be seen in all patients with similar stage tumours, or that the gene expression signatures that define these subgroups would be effective biomarkers. However, these are nevertheless very interesting observations, which are supported by our more recent experimental studies. As such, data generated in this project should provide useful insights, which can be used to guide future studies into the relevance of these trends in larger patient cohorts, or to guide further experimental work into the correlation between CMA or ATM reprogramming, tumour development and patient prognosis.

Considering results from our study in light of information emerging from the study of cancer microenvironment interactions in other tissues, several similarities and novel observations have emerged. Firstly, our data has provided the first information, to suggest that both gastric CAMs and ATMs undergo metabolic reprogramming relative to gastric ANMs. Our finding generally support the principle of the reverse Warburg effect, proposed by Lisanti et al from work performed on Breast tumours (Pavlides et al., 2009b).

The classical 'Reverse Warburg effect' describes a process whereby 'activated' cancer associated fibroblasts carry out aerobic gylcolysis, producing lactate and pyruvate, which is then transported to nearby cancer cells, via membrane transporters (MCT4 & MCT1), where they then feed into the TCA cycle for oxidative phosphorylation (Pavlides et al., 2010b; Pavlides et al., 2009b). In comparison to this model, we did not identify any alterations in glycolytic pathways that would result in an end product of pyruvate or lactate. However, we found that CAMs have increase expression of the fatty acid transporter SCL27A1 and up-regulation of fatty

acid β-Oxidation pathways. In comparison the SCL27A1 transporter was not up regulated in ATM and fatty acid β-Oxidation appeared reduced within ATM vs. ANM data. In addition, gastric CAMs showed evidence of the conversion of acetyl-CoA into transportable ketone bodies, such as acetoacetate, which has been shown to 'fuel' tumour cell metabolism in other studies (Pavlides et al., 2010a). Also, the ketone body 3-hydroxy-butyrate is known to increased cancer cell growth and act as a chemo-attractant, which increases cancer cell migration (Bonuccelli et al., 2010a). Our observations on the changes seen in gastric CAMs are also consistent with recent evidence that mitochondrial respiration in increased in cancer cells relative to adjacent stromal cells (Bonuccelli et al., 2010a). Overall, increased production of ketone bodies and lactate are shown to correlate with poor clinical outcome. As increases in acetyl-CoA and ketone bodies have both been reported to enhance epigenetic remodelling, leading to changes in gene expression (Martinez-Outschoorn et al., 2011) it would be interesting to investigate the effects that elevated lactate and ketone body levels may have on gene expression in gastric CAMs, ATMs or ANMs.

The MCT4 transporter is up regulated in breast cancer-associated fibroblasts displaying the 'Reverse Warburg' phenotype (Pavlides et al., 2009b); In this case, increased MCT4 expression is observed in CAMs with 'bad' patient survival scores. Therefore, MCT inhibitors were suggested as possible anti-cancer therapies (Bonuccelli et al., 2010a). Due to its strong links with the 'Reverse Warburg effect', we looked into the MCT4 transporter family and found they can also transport ketone bodies out of the cell. In our studies we find that in gastric myofibroblasts

MCT4 is progressively up-regulated (CAMs >ATMs>ANMs) based on proximity to the site of the tumour (increased 5.5 fold in CAMs and 3.5 fold in ATMs).

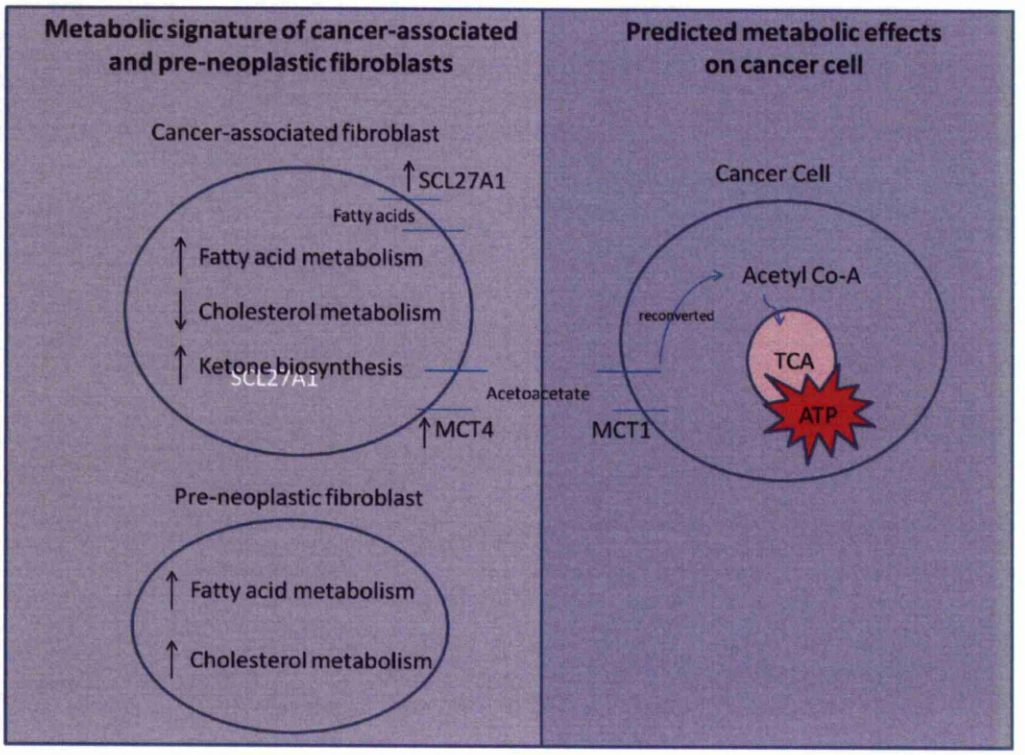



**Figure 6.1** Model of metabolic signature observed in gastric CAMs and ATMs. IN this alternative 'reverse Warburg effect' gastric cancer myofibroblasts up regulate fatty acid metabolism resulting in increased production of ketone bodies, which are then transported to 'fuel' the cancer cell. ATMs (pre-neoplastic fibroblasts) do not exhibit this phenotype. This may be due to reduced exposure to cancer cell induced 'metabolic reprogramming'. Cholesterol and fatty acid metabolism are also up-regulated in ATMs, however in this case cholesterol biosynthesis appears to be routed into increased fatty acid storage as opposed to energy metabolism as observed in CAMs.

Our next aim is to understand whether hypoxia or oxidative stress contributes to 'metabolic reprogramming' of CAMs or ATMs. Several studies have shown that hypoxia is caused by cancer cells activating the transcription factor HIF within cancer-associated fibroblasts. HIF binding to hypoxia response element (HRE) sites within the MCT4 promoter region directly causes an increase in MCT4 expression.

The abolishment of HIF or use of an oxidative stress inhibitor prevents hypoxia-induced production of MCT4 (Robey et al., 2005; Ullah et al., 2006; Whitaker-Menezes et al., 2011). Also, deletion of the oxidative defence gene, Jun D, in cancer stroma of mice resulted in an increased activation of cancer associated fibroblasts and increased metastasis and angiogenesis (Toullec et al., 2010). MCT4 is directly related to 'aerobic glycolysis' and links in with other studies showing that cancer cells induce oxidative stress in fibroblasts by inducing aerobic glycolysis (Martinez-Outschoorn et al., 2010). In terms of our studies on gastric myofibroblasts we now need to investigate whether oxidative stress can trigger the increase in 'fatty acid metabolism' observed in our study.

Overall, our data provides increased insight into the specificity and range of molecular changes that are altered in gastric CAMs and starts to reveal mechanisms by which these phenotypes may occur. It is also important to note that most of the work performed on the breast tumour microenvironment (which led to the development of the reverse Warburg effect, cancer induced pseudo-hypoxia and cancer induced mitophagy) were performed on fibroblast cell lines, not activated tumour-conditioned myofibroblasts. As such, there may be significant differences in the way that these cells respond to the effects of tumour cell conditioning in comparison to the tumour-conditioned CAMs and ATMs used in our studies.

Our preliminary studies focused on identifying the differences between myofibroblasts derived from different regions of the tumour, in all stages of cancer, however in our later studies (chapter 4 and 5) an attempt was made to reveal

differences occurring between myofibroblasts isolated from 'good' and 'bad' patient prognosis groups.

Biological pathways that appear to be important in 'bad' patient prognosis CAMs include cell-cycle pathways, suggesting an overall decrease in CAM proliferation. Interestingly, three separate over-represented pathways are involved in DNA damage detection and down-regulation of caretaker genes, such as Brca1. No such signature arises within the 'good' tumour classification dataset. Also, mitochondrial ketone body biosynthesis is over-represented in CAMs derived from bad prognosis patients. In 'good' prognosis CAMs, we observed increased expression of a range of anti-oxidant enzymes, which could represent a protective mechanism, to retard tumour development work is now on-going to assess the relative levels of these enzymes in all CAM and ATM cell lines.

In our studies we observed a strikingly clear trend in that the variance of gene expression profiles for myofibroblast cells derived from 'good' or 'bad' patient prognosis samples displayed very different patterns, in that variance of gene expression was found to be much larger in myofibroblasts derived from patients with 'bad' prognosis scores (based on the stage of tumour development). To our knowledge this is the first time that such simple criteria has been shown to resolve patients into subgroups that directly match tumour stage scoring. Following this observation, we wished to identify genes that had statistically different expression profiles between 'good' and 'bad' patients; as this analysis could in theory provide clues towards the reason of such differences. Whilst, gene signatures characteristic of each patient prognosis group were identified in our study this analysis is clearly

underpowered, due to the low number of sample available from each prognostic group. However, these are the first indications that this may be possible for gastric cancer samples, and this information can be used to systematically test these trends in future larger studies, involving greater numbers of patient samples.

Having performed a retrospective analysis of different normalisation methods it is feasible that repeating this study on RMA normalised data could provide larger and more informative prognosis group specific signatures. As yet this information does not exist for CAMs derived from other tumour types. Finally, it is interesting to note that prognosis groups could not be identified by correlation analysis, using ATM/ANM profiles. This suggests that it is the type, or extent of reprogramming by cancer cells that confers prognosis related effects in myofibroblasts within the immediate vicinity of the tumour.

It is highly likely that cancer induced reprogramming of CAMs and ATMs is in part mediated by imposed epigenetic remodelling. Indeed, this is most probably why CAMs and ATMs retain programmed characteristics even when isolated from the tissue niche. With this in mind it is interesting to note that a larger number of genes involved in epigenetic regulation are differentially regulated in the 'bad' prognosis CAM sub-group in comparison to the 'good' prognosis CAM sub-group. Previous work in the field has shown that CAMs are more hypo-methylated that ANMs, however, no information currently exists to relate these changes to prognosis or the stage of tumour development. On-going work in the Sanderson lab has now shown that CAMs, ATMs and ANMs respond differently to drugs that modify epigenetic regulation. Also, it appears that exposure to different forms of

oxidative stress induce different changes in epigenetic status, which is linked to the ability of CAMs or ATMs to induce cancer cell migration or proliferation. Again these observations demonstrate that predictions made from our microarray analysis can provide meaningful leads to drive future experimental research.

In conclusion, there are clear limitations to the observations and predictions reported in this study, which are largely due to the restricted number of samples available for analysis. However, several common trends identified in myofibroblasts studied in other tissues were also detected in our study, suggesting that our data is meaningful and informative. In addition, we were able to make new predictions as to the changes, which may contribute to the differential properties of gastric CAMs or ATMs. Finally, the true value of the data and predictions generated in this study will be to provide information to guide future experimental studies into the molecular mechanisms that drive the development of gastric tumours, and secondly, to suggest candidates, which could in future be explored as potential biomarkers. Given that no such markers or effective treatments exist for gastric cancer, we hope that this data will be a useful contribution to the field.

## Reference List


Adams,J.M. and Cory,S. (2007) The Bcl-2 apoptotic switch in cancer development and therapy. *Oncogene*, 26, 1324-1337.

Akre,K. et al. (2001) Aspirin and risk for gastric cancer: a population-based case-control study in Sweden. *Br. J. Cancer*, 84, 965-968.

Alon,Y. et al. (1987) Association in the Expression of Kirsten-Ras Oncogene and the Major Histocompatibility Complex Class-I Antigens in Fibrosarcoma Tumor-Cell Variants Exhibiting Different Metastatic Capabilities. *Cancer Research*, 47, 2553-2557.

Atherton,J.C. (2006) The pathogenesis of Helicobacter pylori-induced gastro-duodenal diseases. *Annu. Rev. Pathol.*, 1, 63-96.

Athippozhy,A. et al. (2011) Differential gene expression in liver and small intestine from lactating rats compared to age-matched virgin controls detects increased mRNA of cholesterol biosynthetic genes. *Bmc Genomics*, 12.

Aune,G. et al. (2011) Increased circulating hepatocyte growth factor (HGF): A marker of epithelial ovarian cancer and an indicator of poor prognosis. *Gynecol. Oncol.*, 121, 402-406.

Awan,A. et al. (2007) Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network. *IET. Syst. Biol.*, 1, 292-297.

Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *Bmc Bioinformatics*, 4, 2.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5, 101-113.

Barclay,W.W. et al. (2005) A system for studying epithelial-stromal interactions reveals distinct inductive abilities of stromal cells from benign prostatic hyperplasia and prostate cancer. *Endocrinology*, 146, 13-18.

Barron,D.A. and Rowley,D.R. (2012) The reactive stroma microenvironment and prostate cancer progression. *Endocr. Relat Cancer*, 19, R187-R204.

Bavelas,A. (1948) A Mathematical Model for Group Structures. *Applied Anthropology*, 7, A16-A30.

Beales,I.L. and Calam,J. (1998) Interleukin 1 beta and tumour necrosis factor alpha inhibit acid secretion in cultured rabbit parietal cells by multiple pathways. *Gut*, 42, 227-234.

Beltinger,J. et al. (1999) Human colonic subepithelial myofibroblasts modulate transepithelial resistance and secretory response. *Am. J. Physiol*, 277, C271-C279.

Berdasco,M. and Esteller,M. (2010) Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Dev. Cell*, 19, 698-711.

Bernfield,M. et al. (1999) Functions of cell surface heparan sulfate proteoglycans. *Annu. Rev. Biochem.*, 68, 729-777.

Bernstein,A.M. et al. (2007) Urokinase receptor cleavage: a crucial step in fibroblast-to-myofibroblast differentiation. *Mol. Biol. Cell*, 18, 2716-2727.

Bhardwaj,N. (2007) Harnessing the immune system to treat cancer. *J. Clin. Invest*, 117, 1130-1136.

Bhowmick,N.A. et al. (2004a) TGF-beta signaling in fibroblasts modulates the oncogenic potential of adjacent epithelia. *Science*, 303, 848-851.

Bhowmick,N.A., Neilson,E.G. and Moses,H.L. (2004b) Stromal fibroblasts in cancer initiation and progression. *Nature*, 432, 332-337.

Bianchini,F. et al. (2012) 22 : 6n-3 DHA inhibits differentiation of prostate fibroblasts into myofibroblasts and tumorigenesis. *Br. J. Nutr.*, 108, 2129-2137.

Billiau,A. (1987) Interferons and inflammation. *J. Interferon Res.*, 7, 559-567.

Black,A.R., Black,J.D. and Azizkhan-Clifford,J. (2001) Sp1 and kruppel-like factor family of transcription factors in cell growth regulation and cancer. *J. Cell Physiol*, 188, 143-160.

Bonner,J.C. (2004) Regulation of PDGF and its receptors in fibrotic diseases. *Cytokine Growth Factor Rev.*, 15, 255-273.

Bonuccelli,G. et al. (2010a) Ketones and lactate "fuel" tumor growth and metastasis: Evidence that epithelial cancer cells use oxidative mitochondrial metabolism. *Cell Cycle*, 9, 3506-3514.

Breitkreutz,B.J. et al. (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.*, 36, D637-D640.

Broes,A. et al. (1988) Phenotypic and genotypic characterization of enterotoxigenic Escherichia coli serotype O8:KX105 and

O8:K"2829" strains isolated from piglets with diarrhea. *J. Clin. Microbiol.*, 26, 2402-2409.

Cahill,G.F., Jr. and Veech,R.L. (2003) Ketoacids? Good medicine? *Trans. Am. Clin. Climatol. Assoc.*, 114, 149-161.

Capdeville,R. et al. (2002) Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nat. Rev. Drug Discov.*, 1, 493-502.

Carmeliet,P. et al. (1997) Urokinase-generated plasmin activates matrix metalloproteinases during aneurysm formation. *Nat. Genet.*, 17, 439-444.

Casey,T. et al. (2009) Molecular signatures suggest a major role for stromal cells in development of invasive breast cancer. *Breast Cancer Res. Treat.*, 114, 47-62.

Casey,T.M. et al. (2008) Cancer associated fibroblasts stimulated by transforming growth factor beta1 (TGF-beta 1) increase invasion rate of tumor cells: a population study. *Breast Cancer Res. Treat.*, 110, 39-49.

Ceol,A. et al. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, 38, D532-D539.

Cerami,E. et al. (2010) Automated Network Analysis Identifies Core Pathways in Glioblastoma. *Plos One*, 5.

Chauhan,H. et al. (2003a) There is more than one kind of myofibroblast: analysis of CD34 expression in benign, in situ, and invasive breast lesions. *J. Clin. Pathol.*, 56, 271-276.

Chen,H. et al. (2009) TGF-beta induces fibroblast activation protein expression; fibroblast activation protein expression increases the proliferation, adhesion, and migration of HO-8910PM [corrected]. *Exp. Mol. Pathol.*, 87, 189-194.

Chen,L.W. et al. (2003) The two faces of IKK and NF-kappaB inhibition: prevention of systemic inflammation but increased local injury following intestinal ischemia-reperfusion. *Nat. Med.*, 9, 575-581.

Cheng,N. et al. (2005) Loss of TGF-beta type II receptor in fibroblasts promotes mammary carcinoma growth and invasion through upregulation of TGF-alpha-, MSP- and HGF-mediated signaling networks. *Oncogene*, 24, 5053-5068.

Chernicky,C.L. et al. (2005) Tissue-type plasminogen activator is upregulated in metastatic breast cancer cells exposed to insulin-like growth factor-I. *Clin. Breast Cancer*, 6, 340-348.

Christofk,H.R. et al. (2008) The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature*, 452, 230-U74.

Chuang,H.Y. et al. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3, 140.

Cline,M.S. et al. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, 2, 2366-2382.

Cooper,C.R., Chay,C.H. and Pienta,K.J. (2002) The role of alpha(v)beta(3) in prostate cancer progression. *Neoplasia.*, 4, 191-194.

Correa,P., Piazuelo,M.B. and Camargo,M.C. (2006) Etiopathogenesis of gastric cancer. *Scand. J. Surg.*, 95, 218-224.

Cui,Q. et al. (2007) A map of human cancer signaling. *Mol. Syst. Biol.*, 3, 152.

Davies,M.A. and Samuels,Y. (2010) Analysis of the genome to personalize therapy for melanoma. *Oncogene*, 29, 5545-5555.

de Jong,J.S. et al. (1998a) Expression of growth factors, growth inhibiting factors, and their receptors in invasive breast cancer. I: An inventory in search of autocrine and paracrine loops. *J. Pathol.*, 184, 44-52.

de,B.A. et al. (2003) Identification and characterization of E2F7, a novel mammalian E2F family member capable of blocking cellular proliferation. *J. Biol. Chem.*, 278, 42041-42049.

de-Assis,E.M. et al. (2012) Stromal myofibroblasts in oral leukoplakia and oral squamous cell carcinoma. *Med. Oral Patol. Oral Cir. Bucal.*, 17, e733-e738.

Dolberg,D.S. et al. (1985) Wounding and its role in RSV-mediated tumor formation. *Science*, 230, 676-678.

Du,K. and Montminy,M. (1998) CREB is a regulatory target for the protein kinase Akt/PKB. *J. Biol. Chem.*, 273, 32377-32379.

Dublin,E. et al. (2000a) Immunohistochemical expression of uPA, uPAR, and PAI-1 in breast carcinoma. Fibroblastic expression has strong associations with tumor pathology. *Am. J. Pathol.*, 157, 1219-1227.

Duncan,J.A., Reeves,J.R. and Cooke,T.G. (1998) BRCA1 and BRCA2 proteins: roles in health and disease. *Mol. Pathol.*, 51, 237-247.

Durant,S.T. and Nickoloff,J.A. (2005) Good timing in the cell cycle for precise DNA repair by BRCA1. *Cell Cycle*, 4, 1216-1222.

Durinck,S. et al. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21, 3439-3440.

Dvorak,H.F. (1986) Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing. *N. Engl. J. Med.*, 315, 1650-1659.

Fantin,V.R., St-Pierre,J. and Leder,P. (2006) Attenuation of LDH-A expression uncovers a link between glycolysis, mitochondrial physiology, and tumor maintenance. *Cancer Cell*, 9, 425-434.

Fielding.J,J.S.R.S.R.H.P.M.T.M.M.S.E.V.C.O.T.-U.C.H. (2000) A Randomized Double-Blind Placebo-Controlled Study of Marimastat in Patients with Inoperable Gastric Adenocarcinoma. Proc Am Soc Clin Oncol .

Finak,G. et al. (2008) Stromal gene expression predicts clinical outcome in breast cancer. *Nat. Med.*, 14, 518-527.

Folkman,J. et al. (1971) Isolation of a tumor factor responsible for angiogenesis. *J. Exp. Med.*, 133, 275-288.

Freeman,L.C. (1977) Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40, 35-41.

Fuchs,C.S. and Mayer,R.J. (1995) Gastric carcinoma. *N. Engl. J. Med.*, 333, 32-41.

Fukushima,N. et al. (2004) Characterization of gene expression in mucinous cystic neoplasms of the pancreas using oligonucleotide microarrays. *Oncogene*, 23, 9042-9051.

Fuyuhiro,Y. et al. (2010a) Clinical significance of vimentin-positive gastric cancer cells. *Anticancer Res.*, 30, 5239-5243.

Fuyuhiro,Y. et al. (2010b) Myofibroblasts are associated with the progression of scirrhous gastric carcinoma. *Exp. Ther. Med.*, 1, 547-551.

Gabbiani,G. et al. (1972) Granulation tissue as a contractile organ. A study of structure and function. *J. Exp. Med.*, 135, 719-734.

Gamero,A.M. et al. (2010) STAT2 contributes to promotion of colorectal and skin carcinogenesis. *Cancer Prev. Res. (Phila)*, 3, 495-504.

Gan,Q. et al. (2007) Smooth muscle cells and myofibroblasts use distinct transcriptional mechanisms for smooth muscle alpha-actin expression. *Circ. Res.*, 101, 883-892.

Gentleman,R.C. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5, R80.

Giannoni,E. et al. (2010) Reciprocal activation of prostate cancer cells and cancer-associated fibroblasts stimulates epithelial-mesenchymal transition and cancer stemness. *Cancer Res.*, 70, 6945-6956.

Girvan,M. and Newman,M.E.J. (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 7821-7826.

Goh,K.I. et al. (2007) The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 8685-8690.

Greenman,C. et al. (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, 446, 153-158.

Gruber,A.D. and Pauli,B.U. (1999) Tumorigenicity of human breast cancer is associated with loss of the Ca2+-activated chloride channel CLCA2. *Cancer Res.*, 59, 5488-5491.

Grum-Schwensen,B. et al. (2005) Suppression of tumor development and metastasis formation in mice lacking the S100A4(mts1) gene. *Cancer Res.*, 65, 3772-3780.

Guilford,P. et al. (1998) E-cadherin germline mutations in familial gastric cancer. *Nature*, 392, 402-405.

Guimera,R. and Nunes Amaral,L.A. (2005) Functional cartography of complex metabolic networks. *Nature*, 433, 895-900.

Gunduz,M. et al. (2008) Downregulation of ING3 mRNA expression predicts poor prognosis in head and neck cancer. *Cancer Sci.*, 99, 531-538.

Guo,X. et al. (2008) Stromal fibroblasts activated by tumor cells promote angiogenesis in mouse gastric cancer. *J. Biol. Chem.*, 283, 19864-19871.

Gururajan,M., Posadas,E.M. and Chung,L.W. (2012) Future perspectives of prostate cancer therapy. *Transl. Androl Urol.*, 1, 19-32.

Hall,J.M. et al. (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250, 1684-1689.

Hallinan,J. (2004) Gene duplication and hierarchical modularity in intracellular interaction networks. *Biosystems*, 74, 51-62.

Han,J.D. et al. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430, 88-93.

Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, 100, 57-70.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, 144, 646-674.

Hartwell,L.H. et al. (1999) From molecular to modular cell biology. *Nature*, 402, C47-C52.

Hasnain,S.Z. et al. (2012) The interplay between endoplasmic reticulum stress and inflammation. *Immunol. Cell Biol.*, 90, 260-270.

Hawsawi,N.M. et al. (2008) Breast carcinoma - Associated fibroblasts and their counterparts display neoplastic-specific changes. *Cancer Research*, 68, 2717-2725.

Heiden,M.G.V., Cantley,L.C. and Thompson,C.B. (2009) Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. *Science*, 324, 1029-1033.

Heiss,M.M. et al. (1995) Tumor-associated proteolysis and prognosis: new functional risk factors in gastric cancer defined by the urokinase-type plasminogen activator system. *J. Clin. Oncol.*, 13, 2084-2093.

Hemers,E. et al. (2005) Insulin-like growth factor binding protein-5 is a target of matrix metalloproteinase-7: implications for epithelial-mesenchymal signaling. *Cancer Res.*, 65, 7363-7369.

Heppner,K.J. et al. (1996) Expression of most matrix metalloproteinase family members in breast cancer represents a tumor-induced host response. *Am. J. Pathol.*, 149, 273-282.

Holmberg,C. et al. (2012) Release of TGFbetaig-h3 by gastric myofibroblasts slows tumor growth and is decreased with cancer progression. *Carcinogenesis*, 33, 1553-1562.

Huang,d.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37, 1-13.

Huang,d.W. et al. (2008) DAVID gene ID conversion tool. *Bioinformation.*, 2, 428-430.

Hubbell,E., Liu,W.M. and Mei,R. (2002) Robust estimators for expression analysis. *Bioinformatics*, 18, 1585-1592.

Hussain,S.P., Hofseth,L.J. and Harris,C.C. (2003) Radical causes of cancer. *Nat. Rev. Cancer*, 3, 276-285.

Iacopino,F., Angelucci,C. and Sica,G. (2012) Interactions between normal human fibroblasts and human prostate cancer cells in a co-culture system. *Anticancer Res.*, 32, 1579-1588.

Ieung, W.K. et al. (2002) Review article: intestinal metaplasia and gastric carcinogenesis. *Aliment.Parmacol.Ther.*,16, 1209-1216.

Iglesias-Ara,A. et al. (2010) Accelerated DNA replication in. *Oncogene*, 29, 5579-5590.

Inoue,H. et al. (2002) Prognostic score of gastric cancer determined by cDNA microarray. *Clin. Cancer Res.*, 8, 3475-3479.

Iorio,M.V., Piovan,C. and Croce,C.M. (2010) Interplay between microRNAs and the epigenetic machinery: an intricate network. *Biochim. Biophys. Acta*, 1799, 694-701.

Irizarry,R.A. et al. (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, 31, e15.

Irizarry,R.A. et al. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.*, 4, 249-264.

Iynedjian,P.B. (2009) Molecular physiology of mammalian glucokinase. *Cell Mol. Life Sci.*, 66, 27-42.

Jemal,A. et al. (2010) Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiol. Biomarkers Prev.*, 19, 1893-1907.

Jeong,H. et al. (2001) Lethality and centrality in protein networks. *Nature*, 411, 41-42.

Jonsson,P.F. and Bates,P.A. (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22, 2291-2297.

Jordan,J.D., Landau,E.M. and Iyengar,R. (2000) Signaling networks: the origins of cellular multitasking. *Cell*, 103, 193-200.

Kanazawa,T. et al. (2002a) Poorly differentiated adenocarcinoma and mucinous carcinoma of the colon and rectum show higher rates of loss of heterozygosity and loss of E-cadherin expression due to methylation of promoter region. *Int. J. Cancer*, 102, 225-229.

Kauffmann,A., Gentleman,R. and Huber,W. (2009) arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. *Bioinformatics.*, 25, 415-416.

Kauffmann,A. and Huber,W. (2010) Microarray data quality control improves the detection of differentially expressed genes. *Genomics*, 95, 138-142.

Kerrien,S. et al. (2007) IntAct--open source resource for molecular interaction data. *Nucleic Acids Res.*, 35, D561-D565.

Keshava Prasad,T.S. et al. (2009) Human Protein Reference Database--2009 update. *Nucleic Acids Res.*, 37, D767-D772.

Klymkowsky,M.W. and Savagner,P. (2009) Epithelial-mesenchymal transition: a cancer researcher's conceptual friend and foe. *Am. J. Pathol.*, 174, 1588-1593.

Knox,S.S. (2010) From 'omics' to complex disease: a systems biology approach to gene-environment interactions in cancer. *Cancer Cell Int.*, 10, 11.

Kologlu,M. et al. (2000) A prognostic score for gastric cancer. *American Journal of Surgery*, 179, 521-526.

Komarova,N.L. and Wodarz,D. (2005) Drug resistance in cancer: principles of emergence and prevention. *Proc. Natl. Acad. Sci. U. S. A*, 102, 9714-9719.

Kurose,K. et al. (2002) Frequent somatic mutations in PTEN and TP53 are mutually exclusive in the stroma of breast carcinomas. *Nat. Genet.*, 32, 355-357.

Kusters,J.G., van Vliet,A.H. and Kuipers,E.J. (2006) Pathogenesis of Helicobacter pylori infection. *Clin. Microbiol. Rev.*, 19, 449-490.

Kwei,K.A. et al. (2008) Genomic profiling identifies GATA6 as a candidate oncogene amplified in pancreatobiliary cancer. *PLoS Genet.*, 4, e1000081.

Kwon,T. et al. (2003) Mechanism of histone lysine methyl transfer revealed by the structure of SET7/9-AdoMet. *EMBO J.*, 22, 292-303.

Lai,A. et al. (1999) RBP1 recruits both histone deacetylase-dependent and -independent repression activities to retinoblastoma family proteins. *Mol. Cell Biol.*, 19, 6632-6641.

LAUREN,P. (1965) THE TWO HISTOLOGICAL MAIN TYPES OF GASTRIC CARCINOMA: DIFFUSE AND SO-CALLED INTESTINAL-TYPE CARCINOMA. AN ATTEMPT AT A HISTO-CLINICAL CLASSIFICATION. *Acta Pathol. Microbiol. Scand.*, 64, 31-49.

Lauwers,G.Y. and Riddell,R.H. (1999) Gastric epithelial dysplasia. *Gut*, 45, 784-790.

Lee,J.M. et al. (2006) The epithelial-mesenchymal transition: new insights in signaling, development, and disease. *J. Cell Biol.*, 172, 973-981.

Li,Q. et al. (2012) Obesity and gastric cancer. *Front Biosci.*, 17, 2383-2390.

Li,T. et al. (2009) SH2D4A regulates cell proliferation via the ERalpha/PLC-gamma/PKC pathway. *BMB. Rep.*, 42, 516-522.

Li,Y., Agarwal,P. and Rajagopalan,D. (2008) A global pathway crosstalk network. *Bioinformatics*, 24, 1442-1447.

Liu,Z. et al. (2006) Ebp1 isoforms distinctively regulate cell survival and differentiation. *Proc. Natl. Acad. Sci. U. S. A*, 103, 10917-10922.

Lu,H., Ouyang,W. and Huang,C. (2006) Inflammation, a key event in cancer development. *Mol. Cancer Res.*, 4, 221-233.

Ludwig,S., Klitzsch,A. and Baniahmad,A. (2011) The ING tumor suppressors in cellular senescence and chromatin. *Cell Biosci.*, 1, 25.

Luo,B. et al. (2008) Highly parallel identification of essential genes in cancer cells. *Proc. Natl. Acad. Sci. U. S. A*, 105, 20380-20385.

Lushchak,V.I. (2012) Glutathione homeostasis and functions: potential targets for medical interventions. *J. Amino. Acids*, 2012, 736837.

Macheda,M.L., Rogers,S. and Best,J.D. (2005) Molecular and cellular regulation of glucose transporter (GLUT) proteins in cancer. *J. Cell Physiol*, 202, 654-662.

Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21, 3448-3449.

Magistretti,P.J. (2009) Role of glutamate in neuron-glia metabolic coupling. *Am. J. Clin. Nutr.*, 90, 875S-880S.

Majno,G. (1979) The story of the myofibroblasts. *Am. J. Surg. Pathol.*, 3, 535-542.

Markson,G. et al. (2009) Analysis of the human E2 ubiquitin conjugating enzyme protein interaction network. *Genome Res.*, 19, 1905-1911.

Martens,J.H. et al. (2003) Cascade of distinct histone modifications during collagenase gene activation. *Mol. Cell Biol.*, 23, 1808-1816.

Martinez-Moczygemba,M. et al. (1997) Distinct STAT structure promotes interaction of STAT2 with the p48 subunit of the interferon-alpha-

stimulated transcription factor ISGF3. *J. Biol. Chem.*, 272, 20070-20076.

Martinez-Outschoorn,U.E. et al. (2010) Oxidative stress in cancer associated fibroblasts drives tumor-stroma co-evolution: A new paradigm for understanding tumor metabolism, the field effect and genomic instability in cancer cells. *Cell Cycle*, 9, 3256-3276.

Martinez-Outschoorn,U.E. et al. (2011) Ketones and lactate increase cancer cell "stemness," driving recurrence, metastasis and poor clinical outcome in breast cancer: achieving personalized medicine via Metabolo-Genomics. *Cell Cycle*, 10, 1271-1286.

Matsubara,D. et al. (2009) Subepithelial myofibroblast in lung adenocarcinoma: a histological indicator of excellent prognosis. *Mod. Pathol.*, 22, 776-785.

Matsukuma,A. et al. (1996) A clinicopathological study of asymptomatic gastric cancer. *Br. J. Cancer*, 74, 1647-1650.

Matsumoto,K. et al. (2008) N-Glycan fucosylation of epidermal growth factor receptor modulates receptor activity and sensitivity to epidermal growth factor receptor tyrosine kinase inhibitor. *Cancer Sci.*, 99, 1611-1617.

Matthews,L. et al. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, 37, D619-D622.

Mazure,N.M. et al. (1996) Oncogenic transformation and hypoxia synergistically act to modulate vascular endothelial growth factor expression. *Cancer Res.*, 56, 3436-3440.

McCaig,C. et al. (2006a) The role of matrix metalloproteinase-7 in redefining the gastric microenvironment in response to Helicobacter pylori. *Gastroenterology*, 130, 1754-1763.

McCaig,C. et al. (2006b) The role of matrix metalloproteinase-7 in redefining the gastric microenvironment in response to Helicobacter pylori. *Gastroenterology*, 130, 1754-1763.

Mcgill,R., Tukey,J.W. and Larsen,W.A. (1978) Variations of Box Plots. *American Statistician*, 32, 12-16.

Medina,P.P. et al. (2008) Frequent BRG1/SMARCA4-inactivating mutations in human lung cancer cell lines. *Hum. Mutat.*, 29, 617-622.

Meissauer,A. et al. (1991) Urokinase-type and tissue-type plasminogen activators are essential for in vitro invasion of human melanoma cells. *Exp. Cell Res.*, 192, 453-459.

Micallef,L. et al. (2012a) The myofibroblast, multiple origins for major roles in normal and pathological tissue repair. *Fibrogenesis. Tissue Repair*, 5 Suppl 1, S5.

Michieli,P. et al. (2004) Targeting the tumor and its microenvironment by a dual-function decoy Met receptor. *Cancer Cell*, 6, 61-73.

Micke,P. et al. (2007) In situ identification of genes regulated specifically in fibroblasts of human basal cell carcinoma. *J. Invest Dermatol.*, 127, 1516-1523.

Mishra,P.J. et al. (2008) Carcinoma-associated fibroblast-like differentiation of human mesenchymal stem cells. *Cancer Research*, 68, 4331-4339.

Missiuro,P.V. et al. (2009) Information flow analysis of interactome networks. *PLoS Comput. Biol.*, 5, e1000350.

Mohamed,M.M. and Sloane,B.F. (2006) Cysteine cathepsins: multifunctional enzymes in cancer. *Nat. Rev. Cancer*, 6, 764-775.

Morris,M.E. and Felmlee,M.A. (2008) Overview of the proton-coupled MCT (SLC16A) family of transporters: characterization, function and role in the transport of the drug of abuse gamma-hydroxybutyric acid. *AAPS. J.*, 10, 311-321.

Munger,K. and Howley,P.M. (2002) Human papillomavirus immortalization and transformation functions. *Virus Res.*, 89, 213-228.

Naik,M.U. et al. (2008) Attenuation of junctional adhesion molecule-A is a contributing factor for breast cancer cell invasion. *Cancer Res.*, 68, 2194-2203.

Nakagawa,H. et al. (2004) Role of cancer-associated stromal fibroblasts in metastatic colon cancer to the liver and their expression profiles. *Oncogene*, 23, 7366-7377.

Newman,M.E. (2003) Mixing patterns in networks. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.*, 67, 026126.

Newman,M.E.J. (2001) Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review e*, 6401.

Ng,H.H., Jeppesen,P. and Bird,A. (2000) Active repression of methylated genes by the chromosomal protein MBD1. *Mol. Cell Biol.*, 20, 1394-1406.

Nielsen,B.S. et al. (1996) Messenger RNA for urokinase plasminogen activator is expressed in myofibroblasts adjacent to cancer cells in human breast cancer. *Lab Invest*, 74, 168-177.

Noe,V. et al. (2001) Release of an invasion promoter E-cadherin fragment by matrilysin and stromelysin-1. *J. Cell Sci.*, 114, 111-118.

Noel,A., Jost,M. and Maquoi,E. (2008) Matrix metalloproteinases at cancer tumor-host interface. *Semin. Cell Dev. Biol.*, 19, 52-60.

Norsett,K.G. et al. (2010) Gastrin stimulates expression of plasminogen activator inhibitor (PAI)-1 in gastric epithelial cells. *Am. J. Physiol Gastrointest. Liver Physiol.*

O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, 33, D476-D480.

Oehlert,W. et al. (1979) Gastric mucosal dysplasia: what is its clinical significance? *Front Gastrointest. Res.*, 4, 173-182.

Olumi,A.F. et al. (1999) Carcinoma-associated fibroblasts direct tumor progression of initiated human prostatic epithelium. *Cancer Res.*, 59, 5002-5011.

Paciucci,R. et al. (1998) The plasminogen activator system in pancreas cancer: role of t-PA in the invasive potential in vitro. *Oncogene*, 16, 625-633.

Paget,S. (1989) The distribution of secondary growths in cancer of the breast. 1889. *Cancer Metastasis Rev.*, 8, 98-101.

Parsonnet,J. et al. (1994) Helicobacter pylori infection and gastric lymphoma. *N. Engl. J. Med.*, 330, 1267-1271.

Pauli,B.U. and Lee,C.L. (1988) Organ preference of metastasis. The role of organ-specifically modulated endothelial cells. *Lab Invest*, 58, 379-387.

Pavithran,K. et al. (2002) Adenocarcinoma of the stomach in Fanconi's anemia. *Ann. Hematol.*, 81, 666-667.

Pavlides,S. et al. (2010a) The autophagic tumor stroma model of cancer: Role of oxidative stress and ketone production in fueling tumor cell metabolism. *Cell Cycle*, 9, 3485-3505.

Pavlides,S. et al. (2010b) Transcriptional evidence for the "Reverse Warburg Effect" in human breast cancer tumor stroma and metastasis: similarities with oxidative stress, inflammation, Alzheimer's disease, and "Neuron-Glia Metabolic Coupling". *Aging (Albany. NY)*, 2, 185-199.

Pavlides,S. et al. (2009a) The reverse Warburg effect: aerobic glycolysis in cancer associated fibroblasts and the tumor stroma. *Cell Cycle*, 8, 3984-4001.

Pennacchietti,S. et al. (2003) Hypoxia promotes invasive growth by transcriptional activation of the met protooncogene. *Cancer Cell*, 3, 347-361.

Pepper,S.D. et al. (2007) The utility of MAS5 expression summary and detection call algorithms. *Bmc Bioinformatics*, 8, 273.

Phan,S.H. (2003) Fibroblast phenotypes in pulmonary fibrosis. *Am. J. Respir. Cell Mol. Biol.*, 29, S87-S92.

Piechocki,M.P. (2008) A stable explant culture of HER2/neu invasive carcinoma supported by alpha-Smooth Muscle Actin expressing stromal cells to evaluate therapeutic agents. *BMC Cancer*, 8, 119.

Pietras,K. et al. (2008) Functions of paracrine PDGF signaling in the proangiogenic tumor stroma revealed by pharmacological targeting. *PLoS Med.*, 5, e19.

Polager,S. and Ginsberg,D. (2009) p53 and E2f: partners in life and death. *Nat. Rev. Cancer*, 9, 738-748.

Pompella,A. et al. (2003) The changing faces of glutathione, a cellular protagonist. *Biochem. Pharmacol.*, 66, 1499-1503.

Ponti,C. et al. (2002) Role of CREB transcription factor in c-fos activation in natural killer cells. *Eur. J. Immunol.*, 32, 3358-3365.

Poulogiannis,G., Luo,F. and Arends,M.J. (2012) RAS signalling in the colorectum in health and disease. *Cell Commun. Adhes.*, 19, 1-9.

Powell,D.W. et al. (1999) Myofibroblasts. II. Intestinal subepithelial myofibroblasts. *Am. J. Physiol*, 277, C183-C201.

Quan,P.C. and Burtin,P. (1978) Demonstration of Nonspecific Suppressor Cells in Peripheral Lymphocytes of Cancer-Patients. *Cancer Research*, 38, 288-296.

Raica,M., Cimpean,A.M. and Ribatti,D. (2009) Angiogenesis in pre-malignant conditions. *Eur. J. Cancer*, 45, 1924-1934.

Rakoff-Nahoum,S. (2006) Why cancer and inflammation? *Yale J. Biol. Med.*, 79, 123-130.

Robey,I.F. et al. (2005) Hypoxia-inducible factor-1alpha and the glycolytic phenotype in tumors. *Neoplasia.*, 7, 324-330.

Ronnov-Jessen,L. and Petersen,O.W. (1993) Induction of alpha-smooth muscle actin by transforming growth factor-beta 1 in quiescent human breast gland fibroblasts. Implications for myofibroblast generation in breast neoplasia. *Lab Invest*, 68, 696-707.

Ronnov-Jessen,L. et al. (2002) Differential expression of a chloride intracellular channel gene, CLIC4, in transforming growth factor-beta1-mediated conversion of fibroblasts to myofibroblasts. *Am. J. Pathol.*, 161, 471-480.

Rosell,M., Jones,M.C. and Parker,M.G. (2011) Role of nuclear receptor corepressor RIP140 in metabolic syndrome. *Biochim. Biophys. Acta*, 1812, 919-928.

Samoszuk,M., Tan,J. and Chorn,G. (2005) Clonogenic growth of human breast cancer cells co-cultured in direct contact with serum-activated fibroblasts. *Breast Cancer Res.*, 7, R274-R283.

Sanchez-Tillo,E. et al. (2010) ZEB1 represses E-cadherin and induces an EMT by recruiting the SWI/SNF chromatin-remodeling protein BRG1. *Oncogene*, 29, 3490-3500.

Sandoval,J. and Esteller,M. (2012) Cancer epigenomics: beyond genomics. *Curr. Opin. Genet. Dev.*, 22, 50-55.

Sandstrom,M. et al. (1999) Expression of the proteolytic factors, tPA and uPA, PAI-1 and VEGF during malignant glioma progression. *Int. J. Dev. Neurosci.*, 17, 473-481.

Sato,N., Maehara,N. and Goggins,M. (2004) Gene expression profiling of tumor-stromal interactions between pancreatic cancer cells and stromal fibroblasts. *Cancer Res.*, 64, 6950-6956.

Sayers,E.W. et al. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 37, D5-15.

Schaefer,C.F. et al. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, 37, D674-D679.

Seto,M. et al. (2011) Reduced expression of RAS protein activator like-1 in gastric cancer. *Int. J. Cancer*, 128, 1293-1302.

Shaw,A., Gipp,J. and Bushman,W. (2009) The Sonic Hedgehog pathway stimulates prostate tumor growth by paracrine signaling and recapitulates embryonic gene expression in tumor myofibroblasts. *Oncogene*, 28, 4480-4490.

Sherr,C.J. and McCormick,F. (2002) The RB and p53 pathways in cancer. *Cancer Cell*, 2, 103-112.

Shibata,A. et al. (2001) Histological classification of gastric adenocarcinoma for epidemiological research: concordance between pathologists. *Cancer Epidemiol. Biomarkers Prev.*, 10, 75-78.

Shipley,G.D., Tucker,R.F. and Moses,H.L. (1985) Type beta transforming growth factor/growth inhibitor stimulates entry of monolayer

cultures of AKR-2B cells into S phase after a prolonged prereplicative interval. *Proc. Natl. Acad. Sci. U. S. A*, 82, 4147-4151.

Shu,H. and Li,H.F. (2012) Prognostic effect of stromal myofibroblasts in lung adenocarcinoma. *Neoplasma*, 59, 658-661.

Shureiqi,I. et al. (2007) The transcription factor GATA-6 is overexpressed in vivo and contributes to silencing 15-LOX-1 in vitro in human colon cancer. *FASEB J.*, 21, 743-753.

Sinha,S. and Levine,B. (2008) The autophagy effector Beclin 1: a novel BH3-only protein. *Oncogene*, 27 Suppl 1, S137-S148.

Sipponen,P. et al. (1985) Gastric cancer risk in chronic atrophic gastritis: statistical calculations of cross-sectional data. *Int. J. Cancer*, 35, 173-177.

Skalli,O. et al. (1989) Alpha-smooth muscle actin, a differentiation marker of smooth muscle cells, is present in microfilamentous bundles of pericytes. *J. Histochem. Cytochem.*, 37, 315-321.

Slaughter,D., SOUTHWICK,H.W. and SMEJKAL,W. (1953) Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer*, 6, 963-968.

Smith,H.W. and Marshall,C.J. (2010) Regulation of cell signalling by uPAR. *Nat. Rev. Mol. Cell Biol.*, 11, 23-36.

Steinberg,G.R. and Kemp,B.E. (2009) AMPK in Health and Disease. *Physiol Rev.*, 89, 1025-1078.

Sternlicht,M.D. et al. (1999) The stromal proteinase MMP3/stromelysin-1 promotes mammary carcinogenesis. *Cell*, 98, 137-146.

Stuelten,C.H. et al. (2005) Breast cancer cells induce stromal fibroblasts to express MMP-9 via secretion of TNF-alpha and TGF-beta. *J. Cell Sci.*, 118, 2143-2153.

Sudarsanam,P. and Winston,F. (2000) The Swi/Snf family nucleosome-remodeling complexes and transcriptional control. *Trends Genet.*, 16, 345-351.

Sumimoto,H. et al. (2006) The BRAF-MAPK signaling pathway is essential for cancer-immune evasion in human melanoma cells. *J. Exp. Med.*, 203, 1651-1656.

Sun,Q. et al. (2011) Mammalian target of rapamycin up-regulation of pyruvate kinase isoenzyme type M2 is critical for aerobic glycolysis and tumor growth. *Proc. Natl. Acad. Sci. U. S. A*, 108, 4129-4134.

Surowiak,P. et al. (2006) Stromal myofibroblasts in breast cancer: relations between their occurrence, tumor grade and expression of some tumour markers. *Folia Histochem. Cytobiol.*, 44, 111-116.

Svennevig,J.L. and Svaar,H. (1979) Content and distribution of macrophages and lymphocytes in solid malignant human tumours. *Int. J. Cancer*, 24, 754-758.

Syed,A.S., D'Antonio,M. and Ciccarelli,F.D. (2010) Network of Cancer Genes: a web resource to analyze duplicability, orthology and network properties of cancer genes. *Nucleic Acids Research*, 38, D670-D675.

Symmans,W.F. et al. (2003) Total RNA yield and microarray gene expression profiles from fine-needle aspiration biopsy and core-needle biopsy samples of breast carcinoma. *Cancer*, 97, 2960-2971.

Tan,Y. et al. (1996) FGF and stress regulate CREB and ATF-1 via a pathway involving p38 MAP kinase and MAPKAP kinase-2. *EMBO J.*, 15, 4629-4642.

Taniguchi,H. et al. (2011) Silencing of Kruppel-like factor 2 by the histone methyltransferase EZH2 in human cancer. *Oncogene.*

Thibault,A. et al. (1996) Phase I study of lovastatin, an inhibitor of the mevalonate pathway, in patients with cancer. *Clin. Cancer Res.*, 2, 483-491.

Thien,C.B. and Langdon,W.Y. (2001) Cbl: many adaptations to regulate protein tyrosine kinases. *Nat. Rev. Mol. Cell Biol.*, 2, 294-307.

Toullec,A. et al. (2010) Oxidative stress promotes myofibroblast differentiation and tumour spreading. *EMBO Mol. Med.*, 2, 211-230.

Trujillo,K.A. et al. (2010) Markers of fibrosis and epithelial to mesenchymal transition demonstrate field cancerization in histologically normal tissue adjacent to breast tumors. *Int. J. Cancer.*

Tseng,T.C. et al. (2001) VAM-1: a new member of the MAGUK family binds to human Veli-1 through a conserved domain. *Biochim. Biophys. Acta*, 1518, 249-259.

Tsujino,T. et al. (2007) Stromal myofibroblasts predict disease recurrence for colorectal cancer. *Clinical Cancer Research*, 13, 2082-2090.

Tuxhorn,J.A. et al. (2002a) Reactive stroma in human prostate cancer: induction of myofibroblast phenotype and extracellular matrix remodeling. *Clin. Cancer Res.*, 8, 2912-2923.

Tuxhorn,J.A. et al. (2002b) Inhibition of transforming growth factor-beta activity decreases angiogenesis in a human prostate cancer-reactive stroma xenograft model. *Cancer Res.*, 62, 6021-6025.

Ullah,M.S., Davies,A.J. and Halestrap,A.P. (2006) The plasma membrane lactate transporter MCT4, but not MCT1, is up-regulated by hypoxia through a HIF-1alpha-dependent mechanism. *J. Biol. Chem.*, 281, 9030-9037.

Untergasser,G. et al. (2005) Profiling molecular targets of TGF-beta1 in prostate fibroblast-to-myofibroblast transdifferentiation. *Mech. Ageing Dev.*, 126, 59-69.

Vannella,L., Lahner,E. and Annibale,B. (2012) Risk for gastric neoplasias in patients with chronic atrophic gastritis: a critical reappraisal. *World J. Gastroenterol.*, 18, 1279-1285.

Varela,I. et al. (2011) Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*, 469, 539-542.

Vastrik,I. et al. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, 8, R39.

Vecchione,L. et al. (2009) Novel investigational drugs for gastric cancer. *Expert Opinion on Investigational Drugs*, 18, 945-955.

Vose,B.M. and Moore,M. (1979) Suppressor Cell-Activity of Lymphocytes Infiltrating Human-Lung and Breast-Tumors. *International Journal of Cancer*, 24, 579-585.

Vose,B.M., Vanky,F. and Klein,E. (1977) Human Tumor-Lymphocyte Interaction Invitro .5. Comparison of Reactivity of Tumor-Infiltrating, Blood and Lymph-Node Lymphocytes with Autologous Tumor-Cells. *International Journal of Cancer*, 20, 895-902.

Vuaroqueaux,V. et al. (2007) Low E2F1 transcript levels are a strong determinant of favorable breast cancer outcome. *Breast Cancer Res.*, 9, R33.

Wang,L. et al. (2011) PDGF-induced proliferation of smooth muscular cells is related to the regulation of CREB phosphorylation and Nur77 expression. *J. Huazhong. Univ Sci. Technolog. Med. Sci.*, 31, 169-173.

Wang,T.C. et al. (2000) Synergistic interaction between hypergastrinemia and Helicobacter infection in a mouse model of gastric cancer. *Gastroenterology*, 118, 36-47.

Wang,Z. and Zhang,J. (2007) In search of the biological significance of modular structures in protein networks. *PLoS Comput. Biol.*, 3, e107.

Warburg,O. (1956) Origin of Cancer Cells. *Science*, 123, 309-314.

Weaver,V.M. et al. (1995) The development of a functionally relevant cell culture model of progressive human breast cancer. *Semin. Cancer Biol.*, 6, 175-184.

Wei,J. et al. (2010) Regulation of p53 tumor suppressor by Helicobacter pylori in gastric epithelial cells. *Gastroenterology*, 139, 1333-1343.

Whitaker-Menezes,D. et al. (2011) Evidence for a stromal-epithelial "lactate shuttle" in human tumors: MCT4 is a marker of oxidative stress in cancer-associated fibroblasts. *Cell Cycle*, 10, 1772-1783.

White,E. and DiPaola,R.S. (2009) The double-edged sword of autophagy modulation in cancer. *Clin. Cancer Res.*, 15, 5308-5316.

Williamson,D.H. et al. (1971) Activities of enzymes involved in acetoacetate utilization in adult mammalian tissues. *Biochem. J.*, 121, 41-47.

Willis R. (1961) Pathology of tumours.

Witz,I.P. (2009) The tumor microenvironment: the making of a paradigm. *Cancer Microenviron.*, 2 Suppl 1, 9-17.

Wooster,R. et al. (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science*, 265, 2088-2090.

Wu,G.M., Feng,X. and Stein,L. (2010) A human functional protein interaction network and its application to cancer data analysis. *Genome Biology*, 11.

Wu,K.C. et al. (1999) Phenotypic and functional characterisation of myofibroblasts, macrophages, and lymphocytes migrating out of the human gastric lamina propria following the loss of epithelial cells. *Gut*, 44, 323-330.

Yazhou,C. et al. (2004) Clinicopathological significance of stromal myofibroblasts in invasive ductal carcinoma of the breast. *Tumour. Biol.*, 25, 290-295.

Yip,G.W., Smollich,M. and Gotte,M. (2006) Therapeutic value of glycosaminoglycans in cancer. *Mol. Cancer Ther.*, 5, 2139-2148.

You,W.C. et al. (1998) Helicobacter pylori infection, garlic intake and precancerous lesions in a Chinese population at low risk of gastric cancer. *Int. J. Epidemiol.*, 27, 941-944.

Yu,A. et al. (1977) Concomitant Presence of Tumor-Specific Cytotoxic and Inhibitor Lymphocytes in Patients with Osteogenic-Sarcoma. *New England Journal of Medicine*, 297, 121-127.

Yu,W.H. and Woessner,J.F., Jr. (2000) Heparan sulfate proteoglycans as extracellular docking molecules for matrilysin (matrix metalloproteinase 7). *J. Biol. Chem.*, 275, 4183-4191.

Yurchenco,P.D. (2011) Basement membranes: cell scaffoldings and signaling platforms. *Cold Spring Harb. Perspect. Biol.*, 3.

Zanzoni,A., Soler-Lopez,M. and Aloy,P. (2009) A network medicine approach to human disease. *FEBS Lett.*, 583, 1759-1765.

Zhang,X.F. et al. (1993) Normal and oncogenic p21ras proteins bind to the amino-terminal regulatory domain of c-Raf-1. *Nature*, 364, 308-313.

Zhi,K. et al. (2010) Cancer-associated fibroblasts are positively correlated with metastatic potential of human gastric cancers. *J. Exp. Clin. Cancer Res.*, 29, 66.

Zhong,Q. et al. (2009) Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.*, 5, 321.